The 31st New England Statistics Symposium

New England Statistics Symposium

NESS

University of Connecticut
Storrs, Connecticut
April 21 – 22, 2017

The 2017 New England Statistics Symposium would not be possible without the generous support of our sponsors. Please join us in thanking the following organizations for their contributions to this year's Symposium.

UCONN
COLLEGE OF LIBERAL ARTS AND SCIENCES
DEPARTMENT OF STATISTICS

Liberty Mutual.
INSURANCE

TRAVELERS

Pfizer

IBM

THE HARTFORD

ASA

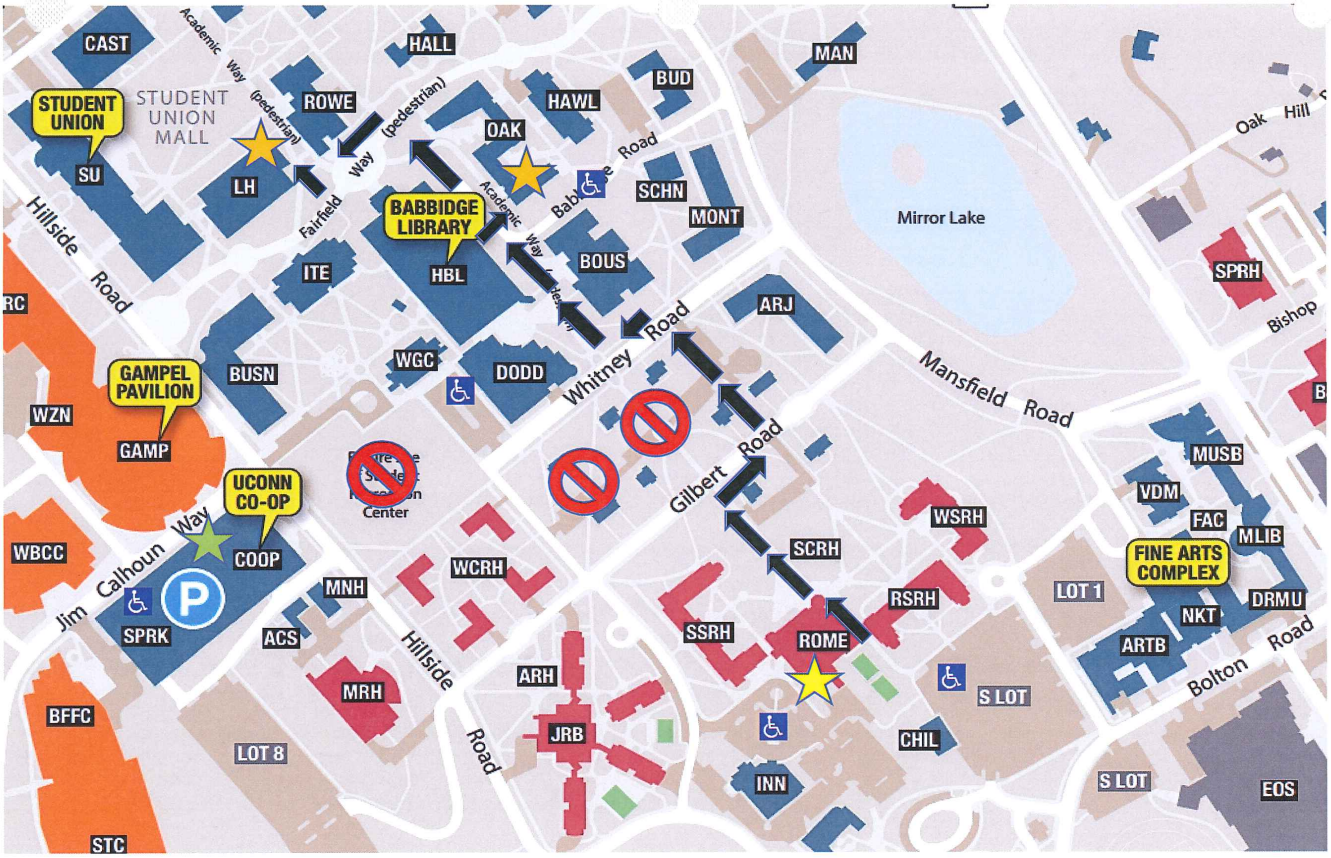
The American Statistical Association
Connecticut Chapter

The first New England Statistics Symposium was held at the University of Connecticut. We are proud to host the event on alternate years and glad to invite other institutions throughout New England to host on alternate years as well.

NESS 2017 Program

Contents

Conference Map



Rome Ballroom (General Sessions and Meals)

Laurel Hall (LH) and Oak Hall (OAK) (Parallel Sessions)

South Parking Garage (Guest Parking)

Handicap Accessible Parking

Campus Construction

Route from Rome Ballroom to Laurel and Oak Halls

Welcoming Remarks

It is my great pleasure to welcome all of you to attend the 31st New England Statistics Symposium. On behalf of the Department of Statistics at the University of Connecticut, I would like to wish all the delegates of the 31st New England Statistics Symposium (NESS) an enjoyable visit to our department and the campus. We are thrilled to have this year two distinguished keynote speakers: Professor Xihong Lin of Harvard School of Public Health and Professor David Madigan of Columbia University. I would also like to thank Professor Douglas Bates, Mr. Jared Huling, and Professor Menggang Yu of University of WisconsinMadison and Professors Kun Chen and Robert Aseltine of University of Connecticut for offering three short courses on Friday, April 21, 2017.

My sincere thanks go to the leadership of Professors Haim Bar, Kun Chen, Jun Yan (Chair) and Yuping Zhang for organizing this symposium. I would like to acknowledge Haim for taking over the chair role since Juns medical leave. I further thank Ms. Megan Petsa, Ms. Tracy Burke, Henry Linder, Hannah Melroy, Gregory Vaughan, and many other graduate student volunteers for their help at each stage of the planning for this symposium. I am confident that the scientific program for this symposium will be of high quality and lead to dynamic exchange of ideas and fruitful interactions between the statistical scientists and graduate students throughout New England and beyond.

I would like to take this opportunity to acknowledge our generous sponsors: College of Liberal Arts & Sciences Deans Office, UConn Department of Statistics, Connecticut Chapter of the American Statistical Association, IBM T.J. Watson Research Center, Liberty Mutual, Pfizer Inc., The Hartford, and Travelers Insurance Company. I would like to thank the IBM T.J. Watson Research Center and Liberty Mutual for sponsoring the Student Awards at the 31st NESS.

Finally, I would like to invite all of you to celebrate the establishment of the New England Statistical Society. Thanks to Xiao-Li Meng, Kun Chen, coordinators and members of the society task force groups for their great effort. Also thanks to the faculty of UConn Department of Statistics for their tremendous support.

Welcome to UConn and Storrs.

With best wishes,

Ming-Hui Chen
Professor and Head, Department of Statistics
University of Connecticut

Dear NESS scholars, Welcome! We are pleased to host the New England Statistics Symposium again this year, especially as this marks the 31st anniversary of your program. I understand that this year?s attendee list boasts more than 230 participants, from over 60 organizations across the nation plus several visitors from overseas. I know that you will all benefit from engaging with your colleagues as you share research and discuss emerging issues in your field. Since its founding in 1939, the College of Liberal Arts and Sciences has been the academic heart of the University of Connecticut. We take seriously the foundations of a liberal education: We teach students to think creatively and analytically; to reason from evidence; to respect the views and experiences of all members of our diverse community; and to continue learning throughout their lives, wherever their professional and personal journeys take them. It is wonderful to hear that the NESS upholds these same values. Best wishes,

Davita Silfen Glasberg
Interim Dean of the College of Liberal Arts and Sciences and Professor of Sociology
University of Connecticut

Welcome to the University of Connecticut! I am delighted that UConn is hosting this conference on its 31st anniversary. We are proud of our Statistics department and honored by this opportunity to host statisticians from across New England and provide a forum for them to discuss emerging issues, share research expertise, and meet old friends.

I would like to particularly thank the Symposium's local hosts and organizers from UConn, Drs. Jun Yan, Haim Bar, Kun Chen, and Yuping Zhang, for their efforts in pulling this meeting together.

Jeremy Teitelbaum
Interim Provost & Executive Vice President for Academic Affairs
University of Connecticut

NESS Organizing Committee

Haim Bar, Kun Chen, Jun Yan, Yuping Zhang

Keynote Speakers

Hypothesis Testing for Weak and Sparse Alternatives With Applications to Whole Genome Data

Dr. Xihong Lin, Harvard University

Massive genetic and genomic data generated using array and sequencing technology present many exciting opportunities as well as challenges in data analysis and result interpretation, e.g., how to develop effective strategies for signal detection using massive genetic and genomic data when signals are weak and sparse. In this talk, I will discuss hypothesis testing for sparse alternatives in analysis of high-dimensional data motivated by gene, pathway/network-based analysis in genome-wide association studies using arrays and sequencing data. I will focus on signal detection when signals are weak and sparse, which is the case in genetic and genomic association studies. I will discuss hypothesis testing for signal detection using variable selection-based penalized likelihood based methods, the Generalized Higher Criticism (GHC) test, the Generalized Berk-Jones test, and the robust omnibus test. I will discuss the challenges in statistical inference in the presence of both between-observation correlation and signal sparsity. The results are illustrated using data from genome-wide association studies and sequencing studies.

Xihong Lin is Chair and Henry Pickering Walcott Professor of Department of Biostatistics and Coordinating Director of the Program of Quantitative Genomics at the Harvard T. H. Chan School of Public Health, and Professor of Statistics of the Faculty of Art and Science of Harvard University.

Dr. Lin's research interests lie in development and application of statistical and computational methods for analysis of massive genetic and genomic, epidemiological, environmental, and medical data. She currently works on whole genome sequencing association studies, genes and environment, analysis of integrated data, and statistical and computational methods for massive health science data.

Dr. Lin received the 2002 Mortimer Spiegelman Award from the American Public Health Association and the 2006 COPSS Presidents' Award. She is an elected fellow of ASA, IMS, and ISI. Dr. Lin received the MERIT Award (R37) (2007–2015), and the Outstanding Investigator Award (OIA) (R35) (2015–2022) from the National Cancer Institute. She is the contacting PI of the Program Project (PO1) on Statistical Informatics in Cancer Research, the Analysis Center of the Genome Sequencing Program of the National Human Genome Research Institute, and the T32 training grant on interdisciplinary training in statistical genetics and computational biology. Dr. Lin was the former Chair of the COPSS (2010–2012) and a former member of the Committee of Applied and Theoretical Statistics (CATS) of the National Academy of Science. She is the former Chair of the new ASA Section of Statistical Genetics and Genomics. She is the former Co-ordinating Editor of Biometrics and the founding co-editor of Statistics in Biosciences and is currently the Associate Editor of Journal of the American Statistical Association. She has served on a large number of statistical society committees, and NIH and NSF review panels.

Honest Learning for the Healthcare System: Large-scale Evidence from Real-world Data

Dr. David Madigan, Columbia University

(joint work with Martijn J. Schuemie, Patrick B. Ryan, George Hripcsak, and Marc A. Suchard)

In practice, our learning healthcare system relies primarily on observational studies generating one effect estimate at a time using customized study designs with unknown operating characteristics and publishing—or not—one estimate at a time. When we investigate the distribution of estimates that this process has produced, we see clear evidence of its shortcomings, including an over-abundance of estimates where the confidence interval does not include one (i.e. statistically significant effects) and strong indicators of publication bias. In essence, published observational research represents unabashed data fishing. We propose a standardized process for performing observational research that can be evaluated, calibrated and applied at scale to generate a more reliable and complete evidence base than previously possible, fostering a truly learning healthcare system. We demonstrate this new paradigm by generating evidence about all pairwise comparisons of treatments for depression for a relevant set of health outcomes using four large US insurance claims databases. In total, we estimate 17,718 hazard ratios, each using a comparative effectiveness study design and propensity score stratification on par with current state-of-the-art, albeit one-off, observational studies. Moreover, the process enables us to employ negative and positive controls to evaluate and calibrate estimates ensuring, for example, that the 95% confidence interval includes the true effect size approximately 95% of time. The result set consistently reflects current established knowledge where known, and its distribution shows no evidence of the faults of the current process. Doctors, regulators, and other medical decision makers can potentially improve patient-care by making well-informed decisions based on this evidence, and every treatment a patient receives becomes the basis for further evidence.

David Madigan is the Executive Vice-President for Arts & Sciences, Dean of the Faculty, and Professor of Statistics at Columbia University in the City of New York. He previously served as Chair of the Department of Statistics at Columbia University (2008–2013), Dean, Physical and Mathematical Sciences, Rutgers University (2005–2007), Director, Institute of Biostatistics, Rutgers University (2003–2004), and Professor, Department of Statistics, Rutgers University (2001-2007). He received his bachelor's degree in Mathematical Sciences (1984, First Class Honours, Gold Medal) and a Ph.D. in Statistics (1990), both from Trinity College Dublin.

Dr. Madigan has over 160 publications in such areas as Bayesian statistics, text mining, Monte Carlo methods, pharmacovigilance and probabilistic graphical models. In recent years, he has focused on statistical methodology for generating reliable evidence from large-scale healthcare data. From 2011 to 2014 he was a member of the FDA's Drug Safety and Risk Management Advisory Committee.

Dr. Madigan is a fellow of the American Association of the Advancement of Science (AAAS), the Institute of Mathematical Statistics (IMS), and the American Statistical Association (ASA), and an elected member of the International Statistical Institute (ISI). He served as Editor-in-Chief of Statistical Science (2008–2010) and Statistical Analysis and Data Mining, the ASA Data Science Journal (2013–2015).

Schedule

Friday, April 21, 2017	
08:30am—05:00pm	NESS short courses at Rome Ballroom
Saturday, April 22, 2017	All activities will be held in Rome Ballroom except where otherwise noted
08:30am—09:15am	Registration & Refreshment & Poster Session
09:15am—09:30am	Welcoming Remarks
09:30am—10:30am	Keynote Presentation:
	David Madigan, Columbia University
10:30am—10:45am	Coffee Break
11:00am—12:45pm	Parallel Invited Sessions **(Laurel / Oak Halls)**
12:45pm—02:00pm	Lunch, Poster Session (continued)
01:00pm—02:00pm	Poster Session (continued)
02:10pm—02:40pm	Special Session: New England Statistical Society
02:40pm—03:40pm	Keynote Presentation:
	Xihong Lin, Harvard University
03:40pm—03:55pm	Coffee Break
04:10pm—05:55pm	Parallel Invited Sessions **(Laurel / Oak Halls)**
05:55pm—06:30pm	Travelers Reception, Student Paper and Poster Awards Ceremony
07:00pm—09:00pm	NESS Dinner (signing up required with limited space; held at **Sichuan Pepper in Vernon**.)

Detailed Program

Morning sessions

1. New Vistas in Statistics with Applications

Organizer: **Aleksey Polunchenko**, Binghamton University
Chair: **Vasanthan Raghavan**, Qualcomm Flarion Technologies, New Jersey

1. **Aleksey Polunchenko**, Binghamton University
   "Asymptotic Exponentiality of the First Exit Time of the Shiryaev-Roberts Diffusion with Constant Positive Drift"
2. **Vasanthan Raghavan**, Qualcomm Flarion Technologies, New Jersey
   "Non-Parametric Approaches to Change Detection"
3. **Zuofeng Shang**, Binghamton University
   "Computationally Efficient Nonparametric Testing"
4. **Emmanuel Yashchin**, IBM
   "Alarm Prioritization in Early Warning Systems"

Oak Hall 235

2. Non-Clinical in Pharmaceutical Industry

Organizer and Chair: **Chi-Hse Teng**, Novartis

1. **Don Bennett**, Pfizer
   "Nonclinical Statistics in Drug Development: In Vitro and in Vivo Examples"
2. **Jerry Lewis**, Biogen
   "Outlook on Outliers"
3. **Ray Liu**, Takeda
   "Big Data, Statistical Innovation and Impact on Drug Development"
4. **Chi-Hse Teng**, Novartis
   "Finding Needles in a Hay Stack - an Approach for a Small-Number-Factor High-Dimensional Data"

Oak Hall 267

3. Space-Time Statistical Solutions at IBM Research

Organizer: **Yasuo Amemiya**, IBM T. J. Watson Research Center
Chair: **Beatriz Etchegaray Garcia**, IBM T. J. Watson Research Center

1. **Julie Novak**, IBM T. J. Watson Research Center
   "Revenue Assessment in Large-Scale Businesses"
2. **Xiao Liu**, IBM T. J. Watson Research Center
   "a Spatio-Temporal Modeling Approach for Weather Radar Image Data"
3. **Rodrigue Ngueyep Tzoumpe**, IBM T. J. Watson Research Center
   "Spatial Segmentation of Spatial-Temporal Lattice Models for Agricultural Management Zoning"
4. **Yasuo Amemiya**, IBM T. J. Watson Research Center
   "Spatio-Temporal Analysis for System Management"

Oak Hall 269

4. Graphical Models, Networks, Regulomes and Multivariate Analysis

Organizer and Chair: **Yuping Zhang**, University of Connecticut

1. **Forrest W. Crawford**, Yale University
   "Causal Inference for Network Epidemics"
2. **Zhengqing Ouyang**, The Jackson Laboratory for Genomic Medicine
   "Understanding Dynamic Regulomes through 3D Genome and Transcriptome Modeling"
3. **Sijian Wang**, University of Wisconsin Madison
4. **Kuang-Yao Lee**, Yale University
   "Learning Causal Networks via Additive Faithfulness"

Oak Hall 268

5. Big Data

Organizer and Chair: **Haim Bar**, University of Connecticut

1. **Jacob Bien**, Cornell University
   "Learning Local Dependence in Ordered Data"
2. **Li Ma**, Duke University
   "Fisher Exact Scanning for Dependency"
3. **Yuwen Gu**, University of Minnesota
   "Penalized Composite Quantile Regression for High-Dimensional Data"

4. **Chihwa Kao**, University of Connecticut
   "Large Dimensional Econometrics and Identification"

Laurel Hall 301

6. Bayesian Applications in High-Dimensional and Multivariate Modeling

Organizer and Chair: **Seongho Song**, University of Cincinnati

1. **Seongho Song**, University of Cincinnati
   "Bayesian Multivariate Gamma-Frailty Cox Model for Clustered Current Status Data"
2. **Xia Wang**, University of Cincinnati
   "Scalable Massive Multivariate Data Modeling"
3. **Gyuhyeong Goh**, Kansas State University
   "Bayesian Variable Selection using Marginal Posterior Consistency"
4. **Jian Zou**, Worcester Polytechnic Institute
   "High Dimensional Dynamic Modeling for Massive Spatio-Temporal Data"

Laurel Hall 308

7. New Advances in Analysis of Complex Data: Heterogeneity and High Dimensions

Organizer and Chair: **Min-ge Xie**, Rutgers University

1. **Dungang Liu**, University of Cincinnati
   "Nonparametric Fusion Learning: Synthesize Inferences from Diverse Sources using Confidence Distribution, Data Depth and Bootstrap"
2. **Dan Yang**, Rutgers University
   "Bilinear Regression with Matrix Covariates in High Dimensions"
3. **Pierre Bellec**, Rutgers University
   "Slope Meets Lasso in Sparse Linear Regression"
4. **Yiyuan She**, Florida State University
   "On Cross-Validation for Sparse Reduced Rank Regression"

Laurel Hall 206

8. Machine Learning and Big Data Analytics

Organizer and Chair: **Jinbo Bi**, University of Connecticut

1. **Sanguthevar Rajasekaran**, University of Connecticut
   "The Closest Pair Problem: Algorithms and Applications"

2. **Renato Polimanti**, Yale University
   "Resources to Investigate the Genetic Architecture of Complex Traits: Large-Scale Datasets and Summary Association Data"
3. **Sheida Nabavi**, University of Connecticut
   "Statistical Machine Learning to Identify Candidate Drivers of Drug Resistance in Cancer"
4. **Michael Kane**, Yale University
   "a First Look at using Human Mobility Data to Assess Community Resilience"

Laurel Hall 306

9. Statistical Approaches in Modeling and Incorporating Dependence

Organizer and Chair: **Ting Zhang**, Boston University

1. **Mengyu Xu**, University of Central Florida
   "Pearson's Chi-Squared Statistics: Approximation Theory and Beyond"
2. **Kun Chen**, University of Connecticut
   "On Large-scale Predictive Modeling of Mixed and Incomplete Outcomes"
3. **Liliya Lavitas**, Boston University
   "Unsupervised Self-Normalized Change-Point Testing for Time Series"
4. **Buddika Peiris**, Worcester Polytechnic Institute
   "Constrained Inference in Regression"

Laurel Hall 309

10. Biopharmaceutical Statistics

Organizer: **Abidemi Adeniji**, EMD Serono
Chair: **Adina Soaita**, Pfizer

1. **Abidemi Adeniji**, EMD Serono
   "Estimation of Discrete Survival Function Through the Modeling of Diagnostic Accuracy for Mismeasured Outcome Data"
2. **Bushi Wang**, Boehringer-Ingelheim
   "How to Evaluate Type Ii Error Rate with Multiple Endpoints"
3. **Joseph C. Cappelleri**, Pfizer
   "Meta-Analysis of Safety Data in Clinical Trials"
4. **Qiqi Deng**, Boehringer Ingelheim
   "Choosing Timing and Boundary for Futility Analysis Based on Cost-Effective Assessment"
5. **Birol Emir**, Pfizer

Laurel Hall 302

11. Extremes

Organizer and Chair: **Richard Davis, Phyllis Wan**, Columbia University

1. **John Nolan**, American University
   "Mvevd: An r Package for Extreme Value Distributions"
2. **Jingjing Zou**, Columbia University
   "Extreme Value Analysis without the Largest Values: What can be Done?"
3. **Karthyek Murthy**, Columbia University
   "Distributionally Robust Extreme Value Analysis"
4. **Tiandong Wang**, Cornell University
   "Asymptotic Normality of in- And Out-Degree Counts in a Preferential Attachment Model"

Laurel Hall 305

12. Feinberg Memorial Session: Bayesian Statistics with Applications

Organizer and Chair: **Dipak Dey**, University of Connecticut

1. **Edoardo Airoldi**, Harvard University
   "Bayesian Methods for Protein Quantification"
2. **Bani Mallick**, Texas A&M University
   "Fast Sampling with Gaussian Scale-Mixture Priors in High Dimensional Regression"
3. **Sudipto Banerjee**, UCLA
   "High-Dimensional Bayesian Geostatistics"

Laurel Hall 307

Afternoon sessions

1. Panel Discussion on Careers in Statistics

Organizer and Chair: **Naitee Ting**, Boehringer Ingelheim Pharmaceuticals, Inc.

1. **Birol Emir**, Pfizer
2. **Chun Wang**, Liberty Mutual Insurance
3. **Yasuo Amemiya**, IBM T. J. Watson Research Center
4. **Minge Xie**, Rutgers University-New Brunswick

Oak Hall 235

2. Statistical Applications in Finance and Insurance

Organizer and Chair: **Guojun Gan**, University of Connecticut

1. **Liang Peng**, Georgia State University
   "Inference for Predictive Regressions"
2. **Fangfang Wang**, University of Connecticut
   "A Common Factor Analysis of Stock Market Trading Activity"
3. **Oleksii Mostovyi**, University of Connecticut
   "Sensitivity Analysis of the Expected Utility Maximization Problem"
4. **Aritra Halder, Shariq Mohammed, Matthew Lamoureux, Brien Aronov**, University of Connecticut
   "Towards Differential Pricing in Auto Insurance via Large-Scale Predictive Modeling: a Partnership Between Travelers and Uconn"

Oak Hall 267

3. Application of Statistical/Predictive Modeling in Health Related Industry

Organizer and Chair: **Nan Shao**, New York Life Insurance

1. **Xiaoyu Jia**, Icahn School of Medicine at Mount Sinai
   "Opportunities and Challenges in Leveraging Results from Analysis of National Cancer Data Base (Ncdb): a Call for Improvement in Quality and Reproducibility"
2. **Zhaonan Sun**, IBM T. J. Watson Research
   "Exploiting Convolutional Neural Network for Risk Prediction with Medical Feature Embedding"
3. **Victoria Gamerman**, Boehringer Ingelheim Pharmaceuticals, Inc.
   "Focusing on Patients: Going Beyond Rcts"
4. **Nan Shao**, New York Life Insurance
   "Statistical Modeling in the Life Insurance Industry"

Oak Hall 268

4. Survival Analysis

Organizer and Chair: **Sy Han Chiou**, Harvard University

1. **Daniel Nevo**, Harvard University
   "Calibration Models for Survival Analysis with Interval-Censored Exposure or Treatment Starting Time"
2. **Bella Vakulenko-Lagun**, Harvard University
   "Cox Regression for Right-Truncated Data"

3. **Jing Qian**, University of Massachusetts
   "Multiple Imputation of Randomly Censored Covariates in Regression Analysis"

4. **Sangwook Kang** , Yonsei University, Korea
   "Accelerated Failure Time Modeling via Nonparametric Infinite Scale Mixtures"

Oak Hall 269

5. Complex Data/Network Modeling

Organizer and Chair: **Yuan Huang**, Department of Biostatistics, Yale University

1. **Yize Zhao**, Weill Cornell Medical College, Cornell University
   "Hierarchical Feature Selection of the Complex Biomedical Data"

2. **Heather Shappell**, Biostatistics, Boston University
   "Methods for Longitudinal Complex Network Analysis in Neuroscience"

3. **Krista Gile**, Math and Statistics, University of Massachusetts
   "Inference from Link-Tracing Network Samples"

4. **Xizhen Cai**, Temple University
   "Variable Selection for Dynamic Citation Networks"

5. **Xuan Bi**, Department of Biostatistics, Yale University
   "Genome-Wide Mediation Analysis of Psychiatric and Cognitive Traits Through Imaging Phenotypes"

Laurel Hall 301

6. Spatial Analysis of Public Health Data

Organizer and Chair: **Beth Ziniti**, Applied Geosolutions LLC

1. **Harrison Quick**, Dornsife School of Public Health, Drexel University
   "Spatiotemporal Trends in Stroke Mortality"

2. **Joshua Warren**, Yale School of Public Health
   "a Bayesian Spatial Kernel Smoothing Method to Estimate Local Vaccine Uptake using Administrative Records"

3. **Gavino Puggioni**, University of Rhode Island
   "Spatiotemporal Analysis of Vector-Borne Disease Risk"

4. **Chanmin Kim**, Harvard T. H. Chan School of Public Health
   "Public Health Impact of Pollutant Emissions"

Laurel Hall 308

7. Network Data Analysis

Organizer and Chair: **Edoardo M. Airoldi**, Harvard University

1. **Jp Onnela**, Harvard University
   "Inference and Model Selection for Mechanistic Network Models"
2. **Vishesh Karwa**, Harvard University
   "Estimating Average Treatment Effects under Interference: Modes of Failure and Solutions"
3. **Xinran Li**, Harvard University
   "Randomization Inference for Peer Effects"

Laurel Hall 206

8. Statistical Approaches to Data Modeling and Analysis

Organizer and Chair: **Erin Conlon**, University of Massachusetts Amherst

1. **Evan Ray**, University of Massachusetts Amherst
   "Feature-Weighted Ensembles for Probabilistic Time-Series Forecasts"
2. **Daeyoung Kim**, University of Massachusetts Amherst
   "Confidence Distribution Sampling and Its Application"
3. **Patrick Flaherty**, University of Massachusetts
   "a Deterministic Global Optimization Method for Variational Inference"
4. **Matthias Steinruecken**, University of Massachusetts Amherst
   "Unraveling the Demographic History of Modern Humans using Full- Genome Sequencing Data"
5. **Zheng Wei**, University of Massachusetts Amherst
   "On Multivariate Asymmetric Dependence using Multivariate Skew-Normal Copula-Based Regression"

Laurel Hall 306

9. Social Networks and Causal Inference

Organizer and Chair: **Daniel Sussman**, Boston University

1. **Daniel Sussman**, Boston University
   "Optimal Unbiased Estimation of Causal Effects under Network Interference"
2. **Alex Volfovsky**, Duke University
   "Causal Inference in the Presence of Networks: Randomization and Observation"
3. **Dean Eckles**, Massachusetts Institute of Technology
   "Estimating Peer Effects in Networks with Peer Encouragement Designs"

4. **Hyunseung Kang**, University of Wisconsin at Madison
   "Peer Encouragement Designs in Causal Inference with Partial Interference and Identification of Local Average Network Effects"

Laurel Hall 309

10. Statistical Innovations in Genomics

Organizer and Chair: **Zhengqing Ouyang**, The Jackson Laboratory for Genomic Medicine

1. **Hongkai Ji**, Johns Hopkins Bloomberg School of Public Health
   "Single-Cell Rna-Seq Analysis by Spanning Trees"
2. **Pei Wang**, Mount Sinai School of Medicine
   "Constructing Tumor-Specific Gene Regulatory Networks Based on Samples with Tumor Purity Heterogeneity"
3. **Yuping Zhang**, University of Connecticut
   "Integrating Diverse Genomic Data to Estimate Multiple Networks"
4. **Kai Wang**, Columbia University
   "Long Read Sequencing to Study Human Genome Variation"

Laurel Hall 302

11. Recent Developments on High-Dimensional Statistics and Regularized Estimation

Organizer and Chair: **Kun Chen**, University of Connecticut

1. **Ethan Fang**, Pennsylvania State University
   "Blessing of Massive Scale: Spatial Graphical Model Estimation with a Total Cardinality Constraint Approach"
2. **Cheng Yong Tang**, Temple University
   "Sufficient Dimension Reduction with Missing Data"
3. **Sahand Nagahban**, Yale University
   "On Approximation Guarantees for Greedy Low Rank Optimization"
4. **Ting Zhang**, Boston University
   "a Thresholding-Based Prewhitened Long-Run Variance Estimator and Its Dependence-Oracle Property"

Laurel Hall 305

12. Subgroup Analysis

Organizer and Chair: **Xiaojing Wang**, University of Connecticut

1. **Yanxun Xu**, Johns Hopkins University
   "a Nonparametric Bayesian Basket Trial Design"
2. **Lynn Lin**, Pennsylvania State University
   "Clustering with Hidden Markov Model on Variable Blocks"
3. **Jared Huling**, University of Wisconsin-Madison
   "Heterogeneity of Intervention Effects and Subgroup Identification Based on Longitudinal Outcomes"
4. **Wai-Ki Yip**, Foundation Medicine, Inc.
   "Stepp Analysis for Continuous, Binary, and Count Outcomes and Other Recent Stepp Development"

Laurel Hall 307

Abstracts of Invited Papers

Morning sessions

1. New Vistas in Statistics with Applications

- **Aleksey Polunchenko**, Binghamton University
  "Asymptotic Exponentiality of the First Exit Time of the Shiryaev-Roberts Diffusion with Constant Positive Drift"
  Aleksey Polunchenko

  We consider the first exit time of a Shiryaev-Roberts diffusion with constant positive drift from the interval $[0, A]$ where $A > 0$. We show that the moment generating function (Laplace transform) of a suitably standardized version of the first exit time converges to that of the unit-mean exponential distribution as $A \to +\infty$. The proof is explicit in that the moment generating function of the first exit time is first expressed analytically and in a closed form, and then the desired limit as $A \to +\infty$ is evaluated directly. The result is of importance in the area of quickest change-point detection, and its discrete-time counterpart has been previously established - although in a different manner - by Pollak and Tartakovsky (2009).

- **Vasanthan Raghavan**, Qualcomm
  "Non-Parametric Approaches to Change Detection"
  Vasanthan Raghavan

  In this talk, we pursue a statistical non-parametric approach for spurt and downfall detection in activity profiles corresponding to real-time applications. While parametric approaches involve learning the underlying parameters, our approach is based on binning the count data of activity to form observation vectors that can be compared with each other. Motivated by a majorization theory framework, these vectors are then transformed via certain functionals and used in spurt classification. While the parametric approaches often result in either a large number of missed detections of real changes or false alarms, the proposed approach is shown to result in a small number of missed detections and false alarms. Further, the non-parametric nature of the approach makes it attractive for ready applications in a practical context. Performance of the proposed approach in a real terrorist monitoring application are presented.

- **Zuofeng Shang**, Binghamton University
  "Computationally Efficient Nonparametric Testing"
  Zuofeng Shang, Meimei Liu, Guang Cheng

  A recent trend of big data problems is to develop computationally efficient inferential methods that embed computational thinking into uncertainty quantification. In this talk I will introduce two new classes of nonparametric testing that scale well with large datasets. One class is based on randomized sketches which can be implemented in one computer, while the other class requires parallel computing. Our theoretical contribution is to characterize the minimal computational cost that is needed to achieve testing optimality. Optimal estimation is a byproduct. The proposed methods are examined by simulated and real datasets.

- **Emmanuel Yashchin**, IBM Research
  "Alarm Prioritization in Early Warning Systems"
  Emmanuel Yashchin

  In complex manufacturing and business operations, early warning systems (EWSs) ensure timely detection of unfavorable trends. Such systems can be deployed so that they act as search engines, analyzing available data at time points that are either pre-specified or determined based on process information. A round of analysis typically encompasses a large number of data streams that are governed by an even larger set of statistical parameters. Careful design of monitoring procedures ensures a low rate of false alarms. To ensure efficient utilization of personnel, it is important that these alarms are properly prioritized. We discuss methods and statistics relevant in the process of alarm prioritization, and their use in the field of integrated circuit manufacturing.

2. Non-clinical in Pharmaceutical Industry

- **Donald Bennett**, Pfizer
  "Nonclinical Statistics in Drug Development: In Vitro and in Vivo Examples"
  Donald Bennett

  An introductory overview of nonclinical statistics critical role in the drug development process will be reinforced by real world examples. Discussion of issues in non-linear models for in vitro assays combined with examples of in vivo experimental designs and analysis will be used to highlight the impact statisticians make in early drug development. We will discuss statistical methods and training needed for successful statistical contribution in the preclinical space. We will also review the expectations of preclinical scientific collaborations and best practices in contrast with the role of clinical statisticians in drug development.

- **Jerry Lewis**, Biogen
  "Outlook on Outliers"
  Jerry Lewis

  Outlier testing has a long and checkered history in the pharmaceutical/biopharmaceutical industry, and has undergone close regulatory scrutiny since the Barr decision (U.S.A. vs. Barr Labs, 1993). This talk will include some basic observations on the philosophy and practice of outlier testing and proceed to the development of a novel outlier test for plate based relative potency dilution bioassays with limited replication.

- **Ray Liu**, Takeda
  "Big Data, Statistical Innovation and Impact on Drug Development"
  Ray Liu, Cong Li, Jacob Zhang

  Drug development is a lengthy process. Speeding up the drug development is a win-win for both patients and pharma companies. Thanks to the advancing technologies that generate data of new modalities with much richer information, the Big Data has seen its adaption into drug development in the past 3-5 years. While Big Data holds the promise to make drug development more efficient, its characteristics of high volume, velocity and variety present unique challenges to existing analytical infrastructures. In this talk I will use real examples to demonstrate how we use statistical innovations within Takeda to maximize the potential of Big Data and shorten the drug development.

- **Chi-Hse Teng**, Novartis
  "Finding Needles in a Hay Stack - an Approach for a Small-Number-Factor High-Dimensional Data"
  Chi-Hse Teng and Azita Ghodssi

  We present a data analysis approach for a project that engineered mRNA constructs to max expression and half-life of a secreted protein. We selected 22 3UTRs, 105 5UTRs and 93 Signal peptides of highly expressed, secreted skeletal muscle proteins to test their relative influences on expression of the protein of interest. Given 22 3UTRs, 105 5UTRs and 93 Signal peptides, an all-encompassing experiment would have required 214,830 constructs to include all combinations. This would have been an overwhelming effort. The screen was limited to an examination of 218 constructs instead. The data analysis effort is to identify 3UTRs, 5UTRs and Signal peptides that might increase the expression of protein. We developed an approach to rank the sequences of each region. The approach delivered sensible results matching the experimental data of some known sequences ranking. It also performed well in the simulation cases.

3. Space-Time Statistical Solutions at IBM Research

- **Julie Novak**, IBM Research
  "Statistical Challenges of Large-Scale Revenue Forecasting"
  Julie Novak, Stefa Etchegaray Garcia, Yasuo Amemiya

  Large-scale businesses need to have a clear vision of how well they expect to perform within all their different units. This information will directly impact managerial decisions that will in turn affect the future health of the company. In this talk, we focus on the statistical challenges that occur when implementing our revenue forecasting methodology on a weekly basis within a large business. We must provide reasonably accurate forecasts for all the geography/division combinations, which have fundamentally different revenue trends and patterns over time. Our method must be robust to oddities, such as typos in the input or unusual behavior in the data. In addition, our forecasts must be stable over weeks, without sacrificing on accuracy. We describe the statistical methods used to maintain an efficient and effective operational solution.

- **Xiao Liu**, IBM Thomas J. Watson Research Center
  "a Spatio-Temporal Modeling Framework for Weather Radar Image Data in Tropical Southeast Asia"
  Xiao Liu, Vik Gopal, Jayant Kalagnanam

  Tropical storms are known to be highly chaotic and extremely difficult to predict. In tropical countries such as Singapore, the official lead time for the warnings of heavy storms is usually between 15 and 45 minutes because weather systems develop quickly and are of very short lifespan. A single thunderstorm cell, for example, typically lives for less than an hour. Weather radar echoes, correlated in both space and time, provide a rich source of information for short-term precipitation nowcasting. Based on a large dataset of 276 tropical storms events, this work investigates a spatio-temporal modeling approach for two-dimensional radar reflectivity (echo) fields. Under a Lagrangian integration scheme, we model the radar reflectivity field by a spatio-temporal conditional autoregressive process with two components. The first component is the dynamic velocity field which determines the motion of the storm, and the second

component governs the growth or decay of the returned radar echoes. The proposed method is demonstrated and compared with existing methods using real radar image data collected from a number of 276 tropical storm events from 2010 to 2011 in Singapore. The advantage of the proposed method is found in modeling small-scale localized convective weather systems, which are the most important type of storm during the Inter-Monsoon Season in Southeast Asia.

- **Rodrigue Ngueyep**, IBM Thomas J. Watson Research Center
  "Spatial Segmentation of Spatial-Temporal Lattice Models for Agricultural Management Zoning"
  Rodrigue Ngueyep, Huijing Jiang, Youngdeok Hwang

  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.

- **Yasuo Amemiya**, IBM T. J. Watson Research Center
  "Spatio-Temporal Analysis for System Management"
  Yasuo Amemiya, Youngdeok Hwang

  IBM has been providing analytics-based solutions to various large-scale problems relevant for business, government, and society. A goal of such a project is to manage a large physical system effectively based on analysis of various measurements taken over space and time. Statistical analysis methods and ideas are essential part of the overall solution development. In particular, new types of spatio-temporal analysis methods are needed. In this talk, some of large system management projects at IBM Research are described, and the development of appropriate spatio-temporal analysis methods is discussed.

4. Graphical Models, Networks, Regulomes and Multivariate Analysis

- **Forrest W. Crawford**, Yale University
  "Causal Inference for Network Epidemics"
  Forrest W. Crawford, Olga Morozova, Xiaoxuan Cai, Ted Cohen

  Estimating the effect of an infectious disease intervention in a single interconnected population is challenging because subjects may transmit infection to others. Preventive treatment (e.g. vaccination) may exert a direct effect on the person who receives it, and may protect others by preventing infection in the recipient and thereby transmission to someone else, or by reducing the recipient's infectiousness when infected. We outline a causal framework for estimating the direct and indirect effects of a vaccine in a single networked population. Epidemiological

assumptions collapse subjects' outcomes over time into cumulative exposures experienced by susceptible subjects; conditional independence assumptions permit identification of the direct effect and partition of the indirect effect into distinct effects on susceptibility and transmissibility. I describe a semi-parametric class of infectious disease regression models motivated by a continuous-time Markov stochastic epidemic process. I describe the consequences of misspecification of the infection model, and discuss approaches to estimation when either the network or time series of infections are incompletely observed.

- **Kuang-Yao Lee**, Yale University
  "Learning Causal Networks via Additive Faithfulness"
  Kuang-Yao Lee, Tianqi Liu, Bing Li, and Hongyu Zhao

  In this work, we propose a statistical model called additively faithful directed acyclic graph (AFDAG) for causal learning from observational data. Our approach is based on additive conditional independence (ACI), a recently proposed three-way statistical relation that shares many similarities with conditional independence but without resorting to multivariate kernels. This distinct feature strikes a balance between a parametric model and a fully nonparametric model, which makes the proposed model attractive to large networks. For graph inference, we develop an estimator for AFDAG based on a linear operator that characterizes ACI, and establish the consistency and convergence rates of our estimator. Moreover, we prove the uniform consistency of the estimated DAG under a stronger additive faithfulness condition, which appears to be less restrictive than its linear counterpart. We introduce a modified PC-algorithm to implement the estimating procedures efficiently, so that their complexity is determined by the level of sparseness rather than the dimension of the network. Through simulation studies we show that our method outperforms existing methods when commonly assumed conditions such as Gaussian or Gaussian copula distributions do not hold. Finally, the usefulness of AFDAG formulation is demonstrated through an application to a proteomics data set

5. Big Data

- **Jacob Bien**, Cornell University
  "Learning Local Dependence in Ordered Data"
  Guo Yu, Jacob Bien

  In many applications, data come with a natural ordering. This ordering can often induce local dependence among nearby variables. However, in complex data, the width of this dependence may vary, making simple assumptions such as a constant neighborhood size unrealistic. We propose a framework for learning this local dependence based on estimating the inverse of the Cholesky factor of the covariance matrix. Penalized maximum likelihood estimation of this matrix yields a simple regression interpretation for local dependence in which variables are predicted by their neighbors. Our proposed method involves solving a convex optimization problem that decomposes into independent subproblems that can be solved efficiently in parallel. Our method yields a sparse, symmetric, positive definite estimator of the precision matrix, encoding a Gaussian graphical model. We derive theoretical results not found in existing methods attaining this structure. Empirical results show our method performing favorably compared to existing methods. We apply our method to genomic data to flexibly model linkage disequilibrium.

- **Li Ma**, Duke University
  "Fisher Exact Scanning for Dependency"
  Li Ma, Jialiang Mao

  We introduce a methodcalled Fisher exact scanning (FES)for testing and identifying variable dependency that generalizes Fishers exact test on 2-by-2 contingency tables to R-by-C contingency tables and continuous sample spaces. FES proceeds through scanning over the sample space using windows in the form of 2-by-2 tables of various sizes, and on each window completing a Fishers exact test. Based on a factorization of Fishers multivariate hypergeometric (MHG) likelihood into the product of the univariate hypergeometric likelihoods, we show that there exists a coarse-to-fine, sequential generative representation for the MHG model in the form of a Bayesian network, which in turn implies the mutual independence (up to deviation due to discreteness) among the Fishers exact tests completed under FES. This allows an exact characterization of the joint null distribution of the p-values and gives rise to an effective inference recipe through simple multiple testing procedures such as Sidak and Bonferroni corrections, eliminating the need for resampling. In addition, FES can characterize dependency through reporting significant windows after multiple testing control. The computational complexity of FES scales linearly with the sample size, which along with the avoidance of resampling makes it ideal for analyzing massive data sets. We use extensive numerical studies to illustrate the work of FES and compare it to several state-of-the-art methods for testing dependency in both statistical and computational performance. Finally, we apply FES to analyzing a microbiome data set and further investigate its relationship with other popular dependency metrics in that context.

- **Yuwen Gu**, University of Minnesota
  "Penalized Composite Quantile Regression for High-Dimensional Data"
  Yuwen Gu

  Composite quantile regression (CQR) provides efficient estimation of the coefficients in linear models, regardless of the error distributions. We consider penalized CQR for both variable selection and efficient coefficient estimation in a linear model under ultrahigh dimensionality and possibly heavy-tailed error distribution. Both lasso and folded concave penalties are discussed. An L2 risk bound is derived for the lasso estimator to establish its estimation consistency and strong oracle property of the folded concave penalized CQR is shown for a feasible solution via the LLA algorithm. The nonsmooth nature of the penalized CQR poses great numerical challenges for high-dimensional data. We provide a unified and effective numerical optimization algorithm for computing penalized CQR via ADMM. We demonstrate the superior efficiency of penalized CQR estimator, as compared to the penalized least squares estimator, through simulated data under various error distributions.

6. Bayesian Applications in High-Dimensional and Multivariate Modeling

- **Seongho Song**, University of Cincinnati
  "Bayesian Multivariate Gamma-Frailty Cox Model for Clustered Current Status Data"
  Negar Jaberansari, Dipak K. Dey and Seongho Song

  Biomedical data analysis plays a key role in today's medicine. Multivariate current status data is a common type of Biomedical data which gives rise to two main challenges in data analysis. First, all event times are censored, making censoring times the only indicator of event

occurrence. Second, an unobserved heterogeneity caused by clusters of units or individuals is probable. To address these issues, mixed Cox proportional hazard model with random block frailty has been used. Here, we consider a Bayesian multivariate Gamma-frailty Cox model and augment the likelihood with respect to random frailties and a set of Poisson latent variables. We also introduce a novel MCMC algorithm by employing two diff erent cumulative baseline hazard function structures: a transformed mixture of incomplete Beta distributions and a linear combination of monotone integrated splines. Through several simulations, we show that our methodology achieves competitive results. We also compare the performance of the two baseline hazard functions using model selection criteria such as AIC and DIC. Finally, we apply the model to a bivariate current status cataract dataset and investigate the e ffect of various risk factors on the occurrence of cataracts.

- **Gyuhyeong Goh**, Kansas State University
  "Bayesian Variable Selection using Marginal Posterior Consistency"
  Gyuhyeong Goh, Dipak K. Dey

  Due to recent technological advancements, high-dimensional data are frequently involved in many areas of science. When an extreme large number of possible predictors are under consideration for the data, marginal likelihood estimation provides an effective way to reduce the high-dimensionality. However, the marginal likelihood-based approach ignores simultaneous influence of predictors and often leads to misidentification of relevant predictors. In this paper, we propose a new variable selection procedure for accounting for the joint influence of important predictors. We use marginal posterior distributions to incorporate all possible predictor effects into the variable selection procedure. Some theoretical properties of the proposed method are investigated. A simulation study demonstrates that our Bayesian approach provides better variable selection performance than existing marginal likelihood methods.

7. New Advances in Analysis of Complex Data: Heterogeneity and High Dimensions

- **Dan Yang**, Rutgers University-New Brunswick
  "Bilinear Regression with Matrix Covariates in High Dimensions"
  Dong Wang, Hongtu Zhu, and Haipeng Shen

  Traditional functional linear regression usually takes a one dimensional functional predictor as input and estimates the continuous coefficient function. Modern applications often generate two dimensional covariates, which when observed at grid points are matrices. To avoid inefficiency of the classical method involving estimation of a two dimensional coefficient function, we propose a bilinear regression model and obtain estimates via a smoothness regularization method. The proposed estimator exhibits minimax optimal property for prediction under the framework of Reproducing Kernel Hilbert Space. The merits of the method are further demonstrated by numerical experiments and an application on real imaging data.

- **Yiyuan She**, Florida State University
  "On Cross-Validation for Sparse Reduced Rank Regression"
  Yiyuan She, Hoang Tran

  In high-dimensional data analysis, regularization methods pursuing sparsity and/or low rank have received a lot of attention recently. To provide a proper amount of shrinkage, it is typical to use a grid search and a model comparison criterion to find optimal regularization

parameters. However, we show that fixing the parameters across all folds may result in an inconsistency issue, and it is more appropriate to cross-validate projection-selection patterns to obtain the best coefficient estimate. Our in-sample error studies in jointly sparse and rank-deficient models lead to a new class of information criteria with four scale-free forms to bypass the estimation of noise level. By use of an identity, we propose a novel scale-free calibration to help cross-validation achieve the minimax optimal error rate non-asymptotically. Extensive simulations support the efficacy of the proposed methods.

8. Machine Learning and Big Data Analytics

- **Renato Polimanti**, Yale University
  "Resources to Investigate the Genetic Architecture of Complex Traits: Large-Scale Datasets and Summary Association Data"
  Renato Polimanti

  Complex traits include a wide range of phenotypes from common diseases to physiological characteristics and their predisposition is generally related to the contribution of hundreds to thousands of variants with small effect. Genome-wide association studies (GWAS) are a powerful approach to detect these small-effect loci, providing an unbiased information to dissect the key biological mechanisms of complex traits. However, to be informative, GWAS require massive cohorts and many genomic consortia, including hundred investigators, are continuing to pool together samples to increase GWAS sample size and consequently their statistical power. There is a pervasive genetic correlation among complex traits due to causal relationships and shared molecular mechanisms. Thus, GWAS results can be used to conduct follow-up investigations, such as polygenic risk score analysis, Mendelian randomization, and phenome-wide association studies, to dissect the genetic architecture of additional traits. Different analytic approaches can be used to conduct these follow-up investigations using large datasets with a wide range of phenotypic information and summary association data. In this talk, I will give an overview of GWAS and post-GWAS studies, describe the current challenges and limitations, and discuss the motivations to develop new statistical methods to investigate the genetics of complex traits.

- **Sheida Nabavi**, University of Connecticut
  "Statistical Machine Learning to Identify Candidate Drivers of Drug Resistance in Cancer"
  Sheida Nabavi

  With advances in technologies, huge amounts of multiple types of high-throughput genomics data are available. These data have tremendous potential to identify new and clinically valuable biomarkers to guide the diagnosis, assessment of prognosis, and treatment of complex diseases. Integrating, analyzing, and interpreting big and noisy genomics data to obtain biologically meaningful results, however, remains highly challenging. Utilizing statistical machine learning methods can help to address these issues. To facilitate the identification of a short list of biologically meaningful genes as candidate drivers of anti-cancer drug resistance from an enormous amount of heterogeneous data, we employed statistical machine-learning techniques and integrated genomics datasets. We developed a computational method that integrates gene expression, somatic mutation, and copy number aberration data of sensitive and resistant tumors. In this method, an integrative method based on regression tree and module network analysis is applied to identify potential driver genes. We applied this method

to the ovarian cancer data from the cancer genome atlas. The method yields a short list of aberrant genes that also control the expression of their co-regulated genes. The final result contains biologically relevant genes, such as COL11A1, which has been recently reported as a cis-platinum resistant biomarker for ovarian carcinoma.

- **Michael Kane**, Yale University
  "a First Look at using Human Mobility Data to Assess Community Resilience"
  Michael Kane

  We use cell tower data to measure communities' ability to respond, withstand and recover from adverse situations. There have recently been calls for significant investments in infrastructure to develop and fortify community resilience. However, resilience has been a difficult concept to quantify, and associated studies tend to be ad-hoc or community specific. We will explore the use of aggregate movement of individuals within a community to measure resiliency. The effect of Hurricane Matthew is used as a case study. We will show the effect of the hurricane on human mobility and identify communities that were particularly hard-hit and slow to respond.

9. Statistical Approaches in Modeling and Incorporating Dependence

- **Mengyu Xu**, University of Central Florida
  "Pearson's Chi-Squared Statistics: Approximation Theory and Beyond"
  Mengyu Xu, Danna Zhang, Wei Biao Wu

  We establish a Chi-squared approximation theory for Pearson's Chi-squared statistics by using a high- dimensional central limit theorem for quadratic forms of random vectors. Our high- dimensional central limit theorem or invariance principle is proved under Lyapunov-type conditions that involve a delicate interplay between the dimension p, the sample size n and the moment condition. To obtain cutoff values of our tests, we introduce a plug-in Gaussian multiplier calibration method and normalized consistency, a new matrix convergence criterion. Based on our modified Chi-squared statistic, we propose the concept of adjusted degrees of freedom. We develop a Cramer-von Mises type test for testing distributions of high dimensional data and develop an approximation theory based on our invariance principle.

- **Kun Chen**, University of Connecticut
  "On Large-scale Predictive Modeling of Mixed and Incomplete Outcomes"
  Chongliang Luo, Jian Liang, Dipak Dey, Fei Wang

  Multivariate outcomes together with multivariate features of possibly high dimensionality have been routinely produced from various fields. In many real-world problems, the collected outcomes are of mixed types, including continuous measurements, binary indicators and counts, and the data may also subject to substantial missing values. Regardless of their types, these mixed outcomes are often interrelated, representing diverse views of the same underlying data generation mechanism. As such, an integrative multivariate modeling can be beneficial. We develop a mixed-outcome reduced rank regression, which effectively enables information sharing among all the prediction tasks. Our approach integrates mixed and partially observed outcomes belonging to the exponential dispersion family, by assuming that all the outcomes are associated through a shared low-dimensional subspace spanned by the high-dimensional features. A general regularized estimation criterion is proposed, and a

unified algorithm with convergence guarantee is developed for optimization. We establish non-asymptotic performance bound for the proposed estimators in the context of mixed outcomes from exponential family and under a general sampling scheme of missing. The effectiveness of our approach is demonstrated by simulation studies and an application on predicting health-related outcomes in longitudinal studies of aging. Other strategies for large-scale prediction, including sequential feature extraction and mixture modeling, will also be discussed.

- **Liliya Lavitas**, Boston University
  "Unsupervised Self-Normalized Change-Point Testing for Time Series"
  Ting Zhang, Liliya Lavitas

  We propose a new self-normalized method for testing change points in the time series setting. Self-normalization has been celebrated for its ability to avoid direct estimation of the nuisance asymptotic variance and its flexibility of being generalized to handle quantities other than the mean. However, it was developed and mainly studied for constructing confidence intervals for quantities associated with a stationary time series, and its adaptation to change-point testing can be nontrivial as direct implementation can lead to tests with nonmonotonic power. Compared with existing results on using self-normalization in this direction, the current paper proposes a new self-normalized change-point test that does not require prespecifying the number of total change points and is thus unsupervised. In addition, we propose a new contrast-based approach in generalizing self-normalized statistics to handle quantities other than the mean, which is specifically tailored for change-point testing. Monte Carlo simulations are presented to illustrate the finite-sample performance of the proposed method.

- **Buddika Peiris**, Worcester Polytechnic Institute
  "Constrained Inference in Regression"
  Buddika Peiris.

  Regression analysis constitutes a large portion of the statistical repertoire in applications. In case where such analysis is used for exploratory purposes with no previous knowledge of the structure one would not wish to impose any constraints on the problem. But in many applications we are interested in a simple parametric model to describe the structure of a system with some prior knowledge of the structure. An important example of this occurs when the experimenter has the strong belief that the regression function changes monotonically in some or all of the predictor variables in a region of interest. Especially in autoregressive models, one can assume that some of parameters are positive (or negative). The analyses needed for statistical inference under such constraints are nonstandard. The specific aim of this study is to introduce a technique, which can be used for statistical inferences of a multiple linear regression with some non-standard constraints.

10. Biopharmaceutical Statistics

- **Abidemi Adeniji**, EMD Serono
  "Estimation of Discrete Survival Function Through the Modeling of Diagnostic Accuracy for Mismeasured Outcome Data"
  Hee-Koung Joeng, Abidemi K. Adeniji, Naitee Ting and Ming-Hui Chen

  Standard survival methods are inappropriate for mismeasured outcomes. Previous research has shown that outcome misclassification can bias estimation of the survival function. We

develop methods to accurately estimate the survival function when the diagnostic tool used to measure the outcome of disease is not perfectly sensitive and specific. Since the diagnostic tool used to measure disease outcome is not the gold standard, the true or error-free outcomes are latent, they cannot be observed. Our method uses the negative predictive value (NPV) and the positive predictive values (PPV) of the diagnostic tool to construct a bridge between the error-prone outcomes and the true outcomes. We formulate an exact relationship between the true (latent) survival function and the observed (error-prone) survival function as a formulation of time-varying NPV and PPV. We specify models for the NPV and PPV that depend only on parameters that can be easily estimated from a fraction of the observed data. Furthermore, we conduct an in depth study to accurately estimate the latent survival function based on the assumption that the biology that underlies the disease process follows a stochastic process. We further examine the performance of our method by applying it to the VIRAHEP-C data.

- **Bushi Wang**, Boehringer Ingelheim
  "How to Evaluate Type Ii Error Rate with Multiple Endpoints"
  Bushi Wang; Naitee Ting

  The FDA draft guidance on multiple endpoints in clinical trials (January 2017) pointed out the regulatory concern of the type II error rate inflation with multiple endpoints. Many of the statistical adjustment to control the type I error rate for multiplicity decrease the study power because they lowered the alpha level used for each of the individual endpoint. The use of co-primary endpoints does not require multiplicity adjustment for type I error but will also increase the type II error rate and decrease study power. In this presentation, I provide a few detailed steps on how to evaluate sample size based on the objective of the clinical study and the selected multiplicity adjustment to control type I error. Analytic forms of power for individual endpoint hypothesis can be derived for most commonly seen scenarios. Simulation can be also easily set up. Optimal sample size is possible by fine tune the individual power for each endpoint with different effect size assumptions.

- **Joseph C. Cappelleri**, Pfizer Inc
  "Meta-Analysis of Safety Data in Clinical Trials"
  Joseph C Cappelleri

  Meta-analyses of clinical trial safety data have risen in importance beyond regulatory submissions. During drug development, pharmaceutical sponsors need to recognize safety signals early and adjust the development program accordingly, so as to facilitate the assessment of causality. Once a medicinal product is marketed, sponsors add post-approval clinical trial data to the body of information to help understand existing safety concerns or those that arise from other post-approval data sources, such as spontaneous reports. The situation becomes more involved when interest centers on a network comparison of multiple active treatments. This presentation highlights some of the major issues considered in meta-analysis of safety data such as sparse events, reporting quality, and limited study duration and identifies gaps requiring special attention.

- **Qiqi Deng**, Boehringer Ingelheim
  "Choosing Timing and Boundary for Futility Analysis Based on Cost-Effective Assessment"
  QiQi Deng, Xiaoqi Lu

  When a futility analyses is included in a trial, its important to choose the right timing for the interim analysis as well as an appropriate futility boundary, so that the trial is likely to be stopped when the interim data suggests a reasonable treatment effect dose not likely

exist. This idea is appealing from an ethical point of view since it may reduce the exposure of patients to ineffective treatments, and from a financial point of view since phase III trials are usually the most significant investment in drug development. However, the design may become inefficient if timing and boundary of futility analysis are not chosen carefully. In this presentation, we will use cost-effectiveness analysis to assess the performance of different futility rules, and introduce a graphical tool to guide the selection of design parameters. In addition, we will discuss how prior information/belief of the treatment effect and other factors may influence the futility decision.

11. Extremes

- **John Nolan**, American University
  "Mvevd: An r Package for Extreme Value Distributions"
  Anne-Laure Fougeres, Cecile Mercadier, John Nolan

  We present a new way to estimate multivariate extreme value distributions (MVEVD) from data using max projections. The approach works in any dimension, though computation time increases quickly as dimension increases. The procedure requires tools from computational geometry and multivariate integration techniques. An R package mevd is being developed to implement the method for several semi-parametric classes of MEVDs: discrete angular measure, generalized logistic, piecewise linear angular measures, and Dirichlet mixture models.

- **Jingjing Zou**, Columbia University in the City of New York
  "Extreme Value Analysis without the Largest Values: What can be Done?"
  Jingjing Zou, Richard A. Davis, Gennady Samorodnitsky

  In this paper we are concerned with the analysis of heavy-tailed data when a portion of the extreme values are unavailable. This research was motivated by an analysis of the degree distributions in a large social network. The degree distributions of such networks tend to have power law behavior in the tails. We focus on the Hill estimator, which plays a starring role in heavy-tailed modeling. The Hill estimator for this data exhibited a smooth and increasing "sample path" as a function of the number of upper order statistics used in constructing the estimator. This behavior became more apparent as we artificially removed more of the upper order statistics. Building on this observation, we introduce a new parameterization into the Hill estimator that is a function of δ and θ, that correspond, respectively, to the proportion of extreme values that are unavailable and the proportion of upper order statistics used in the estimation. As a function of (δ, θ), we establish functional convergence of the normalized Hill estimator to a Gaussian random field. An estimation procedure is developed based on the limit theory to estimate the number of missing extremes and extreme value parameters including the tail index and the bias of Hill's estimate. We illustrate how this approach works in both simulations and real data examples.

- **Karthyek Murthy**, Columbia University in the City of New York
  "Distributionally Robust Extreme Value Analysis"
  Jose Blanchet, Karthyek Murthy

  Typical studies in distributional robustness involve computing worst-case bounds for the quantity of interest (such as expected risk, probability of default, etc.) regardless of the probability distribution used, as long as the distribution lies within a prescribed tolerance

(measured in terms of a probabilistic divergence like KL divergence) from a suitable baseline model.

With this practice of computing worst-case bounds over probabilistic distance based neighborhoods gaining popularity, we go beyond the standard choice of KL divergence to study the role of putative model uncertainty in the context of estimation of tail probabilities or quantiles. In particular, we precisely characterise the worst-case extreme value index in order to answer how heavy the tails of neighboring distributions can be?. This study seeks to understand the qualitative properties of probabilistic distance based neighborhoods in order to guide the selection of model ambiguity regions for estimating extreme quantiles.

- **Tiandong Wang**, Cornell University
  "Asymptotic Normality of in- And Out-Degree Counts in a Preferential Attachment Model"
  Tiandong Wang, Sidney Resnick

  Preferential attachment in a directed scale-free graph is an often used paradigm for modeling the evolution of social networks. Social network data is usually given in a format allowing recovery of the number of nodes with in-degree i and out-degree j. Assuming a model with preferential attachment, formal statistical procedures for estimation can be based on such data summaries. Anticipating the statistical need for such node-based methods, we prove asymptotic normality of the node counts. Our approach is based on a martingale construction and a martingale central limit theorem.

Afternoon sessions

2. Statistical Applications in Finance and Insurance

- **Liang Peng**, Georgia State University
  "Inference for Predictive Regressions"
  Liang Peng

  Testing for predictability of asset returns has been a long history in economics and finance. Recently, based on a simple predictive regression, Kostakis, Magdalinos and Stamatogiannis (2015, Review of Financial Studies) derived a Wald type test based on the context of the extended instrumental variable (IVX) methodology for testing predictability of stock returns and Demetrescu (2014) showed that the local power of the standard IVX-based test could be improved in some cases when a lagged predicted variable is added to the predictive regression on purpose, which poses a general important question on whether a lagged predicted variable should be included in the model or not. This paper proposes novel robust procedures for testing both the existence of a lagged predicted variable and the predictability of asset returns in a predictive regression regardless of regressors being stationary or nearly integrated or unit root. A simulation study confirms the good finite sample performance of the proposed tests. We further apply the proposed tests to some real datasets in finance to illustrate their usefulness in practice.

- **Fangfang Wang**, Uconn
  "A Common Factor Analysis of Stock Market Trading Activity"
  Zhuowang Li, F. Wang

In this talk, we will study the intraday trading activity of U.S. stocks from ten sectors using a new class of parameter-driven models for multivariate count time series that may not be stationary. With the model proposed by F. Wang and H. Wang (2016), we formulate the mean process of trading volume and frequency as the product of modulating factors and unobserved stationary processes. The former characterizes the long-run movement in the trading activity, while the latter is responsible for rapid fluctuations and other unknown or unavailable exogenous/endogenous covariates. The unobserved stationary processes evolve independently of the past observed counts, and might interact with each other. The unobserved processes are further modeled as a linear combination of possibly low-dimensional common factors that govern the contemporaneous and serial correlation within and across the observed counts. With the proposed models, we extract the common factors from the intraday trading volume and frequency of 40 liquid stocks in the first quarter of 2012. Dynamic relationship between common factors adjusted by their associated loading and intraday volatility also investigated.

- **Oleksii Mostovyi**, Uconn
  "Sensitivity Analysis of the Expected Utility Maximization Problem with Respect to Model Perturbations"
  Oleksii Mostovyi, Mihai Sirbu

  We study the sensitivity of the expected utility maximization problem in a continuous semimartingale market with respect to small changes in the market price of risk. Assuming that the preferences of a rational economic agent are modeled with a general utility function, we obtain a second-order expansion of the value function, a first-order approximation of the terminal wealth, and construct trading strategies that match the indirect utility function up to the second order. If a risk-tolerance wealth process exists, using it as a numeraire and under an appropriate change of measure, we reduce the approximation problem to a Kunita-Watanabe decomposition. This talk is based on the joint work with Mihai Sirbu.

- **Brien Aronov, Aritra Halder, Matthew Lamoureux and Shariq Mohammed**, University of Connecticut and Travelers Insurance
  "Modelling of Large Insurance Claims and Occurrence Data: a Uconn - Travelers Partnership"
  Brien Aronov, Kun Chen, Dipak Dey, Aritra Halder, Siddhesh Kulkarni, Matthew Lamoureux, Shariq Mohammed, Elizabeth Schifano and Xiaojing Wang

  This joint presentation features the partnership between Travelers Insurance and the Department of Statistics, University of Connecticut, on analyzing big auto insurance claim data to improve spatial risk classification. In our first student project, we explore a spatial variant of the double generalized linear model (DGLM), in which Tweedie distribution, as a special case, is used to model the pure premium, and the spatial correlation is incorporated via Laplacian regularization. The estimated spatial effects are then used to generate risk rankings at the county level. Simulation results and real data analysis showcase the efficacy of the new methods. Besides our recent progress, the challenges we face in large-scale predictive modeling and our future directions will also be discussed. In particular, we focus on collision data and build models for each state separately.

3. Application of Statistical/Predictive Modeling in Health Related Industry

- **Xiaoyu Jia**, Icahn School of Medicine at Mount Sinai
  "Opportunities and Challenges in Leveraging Results from Analysis of National Cancer Data Base (Ncdb): a Call for Improvement in Quality and Reproducibility"
  Xiaoyu Jia, Madhu Mazumdar

  Use of national registry databases for performing comparative effectiveness research is on rise as they present wonderful opportunity for answering questions about the effectiveness of treatments in the adjuvant or neoadjuvant setting and the associations of patient or tumor characteristics with treatment selection and clinical outcomes. Advanced statistical regression models are available for finding answers to these questions. However, lack of analytic code sharing detailing how the data was manipulated, absence of details about modeling techniques and variables used, and in-sufficient validation of modeling present challenges in understanding how the results could be applicable to ones practice. STROBE and RECORD guidelines are published to guide the design and reporting of observational studies (OS), particularly, those based on routinely collected health care data. Despite emerging evidence that use of reporting guidelines improve quality of reporting, many journals have still not adopted these guidelines and even when adopted, have not mandated their use. We focus our attention to published OS based on National Cancer Data Base (NCDB), a commonly used database in oncology research, and Journal of Clinical Oncology (JCO), a high-impact journal, and a recent time frame of Jan 2015 to March of 2017. We checked the 16 publications found to assess how well they followed the 22 criteria specified by STROBE/RECORD guideline. Best-practices especially those recommended by RECORD on code sharing and model validation were followed at low-moderate rate in the range 0-25%. We call for JCO and others to adopt reporting guidelines seriously. Despite availability of large sample size and a rich array of clinical variables, the results based on NCDB will not progress to clinical practice until we could improve the quality and reproducibility of these studies.

- **Zhaonan Sun**, IBM Research
  "Exploiting Convolutional Neural Network for Risk Prediction with Medical Feature Embedding"
  Zhengping Che, Yu Cheng, Zhaonan Sun, Yan Liu

  The widespread availability of electronic health records (EHRs) promises to usher in the era of personalized medicine. However, the problem of extracting useful clinical representations from longitudinal EHR data remains challenging, owing to the heterogeneous, longitudinally irregular, noisy and incomplete nature of such data.

  In this talk, we will focus on the problems of high dimensionality and temporality. We explore deep neural network models with learned medical feature embedding to deal with these issues. Specifically, we use a multi-layer convolutional neural network (CNN) to parameterize the model and is thus able to capture complex non-linear longitudinal evolution of EHRs. To account for high dimensionality, we extended the word2vec model and use the embedded medical features in the CNN model. Experiments on real-world EHR data demonstrate the effectivenes of the proposed method.

- **Victoria Gamerman**, Boehringer-Ingelheim Pharmaceuticals, Inc.
  "Focusing on Patients: Going Beyond Rcts"
  Steven Edelman, Matthew Capehorn, Anne Belton, Susan Down, Aus Alzaid, Friederike

Nagel, Jisoo Lee, William H. Polonsky

Type 2 diabetes (T2D) presents challenges both for physicians, who often have limited time and resources, and for patients, who can experience psychological and behavioural issues. Effective communication between physicians and patients, especially during the early phases of T2D treatment, may lead to improvements in patient self-care and outcomes, which is important considering the clinical benefits associated with achieving good glycaemic control early in the course of T2D.

IntroDia a large cross-national survey in 26 countries has investigated physician-patient communication during early T2D treatment. The survey was designed in partnership with the International Diabetes Federation and a multidisciplinary advisory board. Around 17,000 participants (physicians and patients with T2D) were surveyed using validated scales and novel questionnaires; these assessed physician-patient communication both at diagnosis and at first prescription of additional oral medication, as well as patient-reported outcomes.

Overall, findings from IntroDia suggest that patient-physician communication at diagnosis of T2D and at add-on may be enhanced by physicians using more collaborative and encouraging and fewer discouraging conversation elements, and this may contribute to patients subsequently experiencing greater well-being and managing the disease more effectively.

Methodologies and results from the survey will be highlighted.

4. Survival Analysis

- **Daniel Nevo**, Harvard University
  "Calibration Models for Survival Analysis with Interval-Censored Exposure or Treatment Starting Time"
  Daniel Nevo, Tsuyoshi Hamada, Shuji Ogino and Molin Wang

  We consider the association of a time-dependent binary treatment or exposure with time-to-event under the proportional hazard model. The exposure value is assumed zero at the beginning of the study and may change to one at any time point. The value of the exposure is observed only in certain time points, and thus its exact value is unknown for some participants, in each risk set. We are motivated by the assessment of post colorectal cancer diagnosis aspiring taking and survival. Nave and popular methods are potentially biased, especially when the exposure is measured at a small number of time points. We present a class of calibration models that fit a distribution for the time to exposure starting time. Estimates obtained from these models are then incorporated in the partial likelihood in a natural way. We derive asymptotic theory for these methods. Our methodology allows for inclusion of further baseline covariates affecting the initiation time of the exposure of interest. Certain bias is expected from our methods when the exposure effect is large, and we provide a less-biased alternative using a risk set calibration approach.

- **Bella Vakulenko-Lagun**, Harvard University
  "Cox Regression for Right-Truncated Data"
  Bella Vakulenko-Lagun, Rebecca Betensky, Micha Mandel

  Right-truncated survival data arise when observations are sampled retrospectively and only those who had experienced the event of interest prior to some sampling time are included in a

sample. As a result, the obtained sample is biased, since those who survive longer have lower probability to be selected. If the interest is in the nonparametric estimation of the lifetime distribution from right- truncated data, then this task can be approached by reversing time and transforming the problem of right-truncation into a well-developed problem of estimation under left truncation. However, when the goal is to explain survival by some covariates, it is unclear how to interpret results from the reverse time analysis in terms of the forward time effects of covariates. Other existing methods for the Cox regression under right truncation, although can be used for testing covariate effect, suffer from an identifiability problem in estimation or are computationally intensive. The proposed approach based on the Inverse-Probability-Weighting (IPW) estimating equations does not have an identifiability problem, it works in a forward time so that covariate effects can be interpreted as usual, it performs better than existing methods for both purposes of testing and estimation, and it is easily implemented using standard software. Methods are compared in simulations and through an application to real data.

- **Jing Qian**, University of Massachusetts-Amherst
  "Multiple Imputation of Randomly Censored Covariates in Regression Analysis"
  Folefac Atem, Jing Qian, Jacqueline Maye, Keith Johnson, Rebecca Betensky

  Randomly censored covariates arise frequently in epidemiologic studies. The most commonly used methods, including complete case and single imputation or substitution, suffer from inefficiency and bias, they make strong parametric assumptions and they consider limit of detection censoring only. We employ multiple imputation, in conjunction with semi-parametric modeling of the censored covariate, to overcome these shortcomings and to facilitate robust estimation. We develop a multiple imputation approach for randomly censored covariates within the framework of linear and logistic regression models. We use the nonparametric estimate of the covariate distribution, or the semi-parametric Cox model estimate in the presence of additional covariates in the model. We evaluate this procedure in simulations, and compare its operating characteristics to those from the complete case analysis and a survival regression approach. We apply the procedures to an Alzheimers study of the association between amyloid positivity and maternal age of onset of dementia.

- **Sangwook Kang**, Yonsei University, Korea
  "Accelerated Failure Time Modeling via Nonparametric Infinite Scale Mixtures"
  Byungtae Seo, Sangwook Kang

  A semiparamtric accelerated failure time (AFT) model resembles the usual linear regression model with the response variable being the logarithm of failure times while the random error term is left unspecified. Thus, it is more flexible than parametric AFT models that assume parametric distributions for the random error term. Estimation for model parameters is typically done through a rank-based procedure, in which the intercept term cannot be directly esitmated. This requires a separate estimation procedure for the intercept, which often leads to unstable estimates. For better estimation of the intercept essential in estimating mean failure times or survival functions, we propose to employ a mixture model approach. To leave the model as flexible as possible, we consider nonparametric infinite scale mixtures of normal distributions. An expectation-maximization (EM) method is used to estimate model parameters. Finite sample properties of the proposed estimators are investigated via an extensive stimulation study. The proposed estimators are illustrated using a real data analysis.

5. Complex Data/Network Modeling

- **Xizhen Cai**, Temple University
  "Variable Selection for Dynamic Citation Networks"
  Xizhen Cai, David Hunter

  Recently, survival models like the Cox model are also extended to apply to dynamic network data (Vu et al., 2011b; Perry and Wolfe, 2013), where the observations are dependent. We extend the penalization idea to the Cox model in an egocentric approach to dynamic networks, and select covariates by maximizing the penalized partial likelihood function. Asymptotic properties of both the unpenalized and penalized partial likelihood estimates are developed under certain regularity conditions. We also implement the estimation and test the prediction performance of these estimates in a citation network. Since the covariates are time-varying, the computation cost is high. After variable selection, the model is reduced, which simplifies the calculation for future predictions. Another method to reduce the computational complexity is to use the case-control approximation, in which instead of using all the at-risk nodes in the network, only a subset is sampled to evaluate the partial likelihood function. By using this approximation, the computation time is shortened dramatically, while the prediction performance is still satisfactory in the citation network.

- **Xuan Bi**, Yale University
  "Genome-Wide Mediation Analysis of Psychiatric and Cognitive Traits Through Imaging Phenotypes"
  Xuan Bi, Liuqing Yang, Tengfei Li, Baisong Wang, Hongtu Zhu, Heping Zhang

  Heritability is well documented for psychiatric disorders and cognitive abilities which are, however, complex, involving both genetic and environmental factors. Hence, it remains challenging to discover which and how genetic variations contribute to such complex traits. In this article, we propose to use mediation analysis to bridge this gap, where neuroimaging phenotypes were utilized as intermediate variables. The Philadelphia Neurodevelopmental Cohort was investigated using genome-wide association studies (GWAS) and mediation analyses. Specifically, 951 participants were included with age ranging from 8 to 21 years. Two hundred and four neuroimaging measures were extracted from structural magnetic resonance imaging scans. GWAS were conducted for each measure to evaluate the SNP-based heritability. Furthermore, mediation analyses were employed to understand the mechanisms in which genetic variants have influence on pathological behaviors implicitly through neuroimaging phenotypes. Our analyses found, rs10494561, located within NMNAT2, to be associated with the severity of the prodromal symptoms of psychosis implicitly, mediated through the volume of the left hemisphere of the superior frontal region. Another SNP rs2285351 was found in the IFT122 gene that may be potentially associated with human spatial orientation ability through the area of the left hemisphere of the isthmuscingulate region.

6. Spatial Analysis of Public Health Data

- **Harrison Quick**, Drexel University
  "Spatiotemporal Trends in Stroke Mortality"
  Harrison Quick

  Geographic patterns in stroke mortality have been studied as far back as the 1960s, when

a region of the southeastern United States became known as the "stroke belt" due to its unusually high rates. While stroke mortality rates are known to increase exponentially with age, an investigation of spatiotemporal trends by age group at the county-level is daunting due to the preponderance of small population sizes and/or few stroke events by age group. Our goal here is two-pronged. First and foremost, we harness the power of a complex, nonseparable multivariate space-time model which borrows strength across space, time, and age group to obtain reliable estimates of yearly county-level mortality rates from US counties between 1973 and 2013 for those aged 65+. Second, we outline how the results of this model fit can be used to generate high-quality synthetic data for public use that preserve data confidentiality without sacrificing data utility.

- **Joshua Warren**, Yale University
  "a Spatial Method to Estimate Local Vaccine Uptake using Administrative Records"
  Joshua Warren, Esra Kurum, Daniel Weinberger

  It is necessary to quantify the level of vaccine uptake among a population of interest in order to determine if the introduced vaccine has the desired beneficial impact on human health. A number of data sources and methods are available to obtain this information at aggregated spatial levels for many vaccines. However, obtaining an accurate assessment of uptake at more localized spatial scales can be a difficult task due to limitations of regularly collected administrative data. Vaccine recipients often live in one region while being vaccinated in another, thereby complicating the process of calculating uptake within a region. We introduce a spatial kernel smoothing method in the Bayesian setting that allows for estimation of local vaccine uptake through the combination of administrative and survey data sources. The newly developed method is applied to pneumococcal conjugate vaccine uptake data from Brazil in 2013. Results suggest that the method provides estimates of vaccine uptake at local levels that are in closer agreement to collected survey responses than the standard method that ignores the issue of participant mobility. The method also provides insight into patterns of mobility of vaccine recipients based on the inclusion of region-specific covariates.

- **Gavino Puggioni**, University of Rhode Island
  "Spatiotemporal Analysis of Vector-Borne Disease Risk"
  Janelle Couret, Emily Serman, Ali Akanda, Howard Ginsberg

  This work presents the first comprehensive spatio-temporal analysis that links reported and suspected cases of Dengue (recorded monthly in Puerto Rico from 1990 to 2015 in 76 municipalities) with weather variables collected at 34 stations and land use satellite data. Dengue and Zika are mosquito-borne tropical diseases, reported with increasing rates in the last decade. Early warning systems help in predicting outbreaks and allow public health decision-makers to implement preventive measures. Several factors have been linked to the increase in reported cases: changes in temperature, precipitation, urbanization, and other spatial variables. Several space-time CAR specifications are implemented in a Bayesian framework to assess the relative risk of these factors, as well as to set a predictive framework. The modeling strategy involves a two stage approach to account for the different spatial supports of predictors and response.

- **Chanmin Kim**, Harvard University
  "Public Health Impact of Pollutant Emissions"
  Corwin Zigler, Christine Choirat

  Pollutant emissions from coal burning power plants have been deemed to adversely impact

ambient air quality and public health conditions. Over the last few decades, many air quality control strategies targeting emissions have been adopted at the U.S. power plants. Despite noticeable reduction in emissions and the improvement of air quality since the Clean Air Act (CAA) became the law, the public-health benefits from changes in emissions have not been widely evaluated yet. In terms of the chain of accountability, the link between pollutant emissions and public health conditions with counting for changes in ambient air quality, we provide the first epidemiological assessment of the health effect of specific pollutant emission (SO2) that is mediated through change in the ambient air quality. Especially, we pursue the link from SO2 emissions from coal-fired power plants (intervention) to ambient PM2.5 concentrations (mediator) estimated for each zip code and from ambient PM2.5 to cardiovascular- and respiratory-hospitalization and all causes mortality (outcomes). The main linkage is based on the HYSPLIT model developed by the National Oceanic and Atmospheric Administration (NOAA), which simulates air mass trajectories from coal-fired power plants. To draw causality in the observational data, we use the potential outcomes framework with direct adjustment for confounding variables in the regression model. Then, we use a newly-developed Bayesian nonparametric method to provide flexible models to the observed data in two analyses: principal stratification analysis and mediation analysis. Both analyses are anchored to the same observed data model and used as the means to quantify the effects through two causal pathways: the extent to which SO2 emissions affect public health outcomes that is attributable to changes in ambient PM2.5 and the extent to which SO2 emissions directly affect public health outcomes.

7. Network Data Analysis

- **Jp Onnela**, Harvard University
  "Parameter Inference and Model Selection for Mechanistic Network Models"
  JP Onnela, Antonietta Mira

  There are (at least) two prominent paradigms to the modeling of networks, which might be called the statistical approach and the mechanistic approach. In the statistical approach, one describes a model that specifies the likelihood of observing a given network, i.e., these are probabilistic models of data that happen to arrive in the shape of a network. In the mechanistic approach, one specifies a set of domain-specific mechanistic rules, based on scientific understanding of the problem, that are used to grow or evolve the network over time. Both modeling approaches provide distinct angles and advantages to our understanding of complex systems. I propose to discuss two interrelated topics. First, I will present a framework to identify the essential network properties associated with mechanistic network models. The joint distribution of these properties determines whether it is feasible to model the given networks using a specific statistical framework. Second, I will discuss a general framework for both parameter inference and model selection for mechanistic network models. In the former, a mechanistic network model might incorporate some parameter whose value is not fixed and must therefore be learned from data. In the related problem of model selection, the goal is to learn our degree of belief that any of the candidate mechanistic models was used to generate the network. The proposed approach can quantify the probability that any given model in the collection generated the data, and this approach can therefore be used to select the most likely model from among the collection of candidate models.

- **Vishesh Karwa**, Harvard University

"Estimating Average Treatment Effects under Interference: Modes of Failure and Solutions"
Vishesh Karwa, Edo Airoldi

Estimating average treatment effects in the presence of network interference has recently gained a lot of attention. It is well known that the classical versions of average treatment effects are no longer well defined when there is interference. We will investigate the issues that arise in defining and estimating average causal effects when there is arbitrary interference. We posit models for specifying various forms of interference through the concept of an *exposure* neighborhood and develop a linear non-parametric representation of potential outcomes. Focusing on unbiased estimation (with respect to the randomization distribution), we will study two types of average causal effects. We will examine the pitfalls of ignoring interference and the consequences of using classical designs and estimators for estimating average causal effects. The classical difference-of-means estimators can have arbitrary bias. The nature and source of bias depends on the form of interference, which is unknown in general. On the other hand, Horvitz-Thompson estimators are unbiased as along as the correct weights are used; these weights depend on the form of interference. We also show that the H-T estimators are in-admissible for a large class of designs.

- **Xinran Li**, Harvard University
  "Randomization Inference for Peer Effects"
  Xinran Li, Peng Ding, Qian Lin, Dawei Yang and Jun S. Liu

  Many previous causal inference studies required no interference among units, that is, the potential outcomes of a unit do not depend on the treatments of other units. This no-interference assumption, however, becomes unreasonable when units are partitioned into groups and they interact with other units within groups. In a motivating education example from Peking University, students are admitted either through the college entrance exam (also known as Gaokao), or recommendation (often based on Olympiads in various subjects). Right after entering college, students are randomly assigned to different dorms, each of which hosts four students. Because students within the same dorm live together almost every day and they interact with each other intensively, it is very likely that peer effects exist and the no-interference assumption is violated. More importantly, understanding peer effects among students gives useful guidance for future roommate assignment to improve the overall performances of the students. Methodologically, we define peer effects in terms of potential outcomes, and propose a randomization-based inference framework to study peer effects in general settings with arbitrary numbers of peers and arbitrary numbers of peer types. Our inferential procedure does not require any parametric modeling assumptions on the outcome distributions. Our analysis of the data set from Peking University gives useful practical guidance for policy makers.

8. Statistical Approaches to Data Modeling and Analysis

- **Evan L. Ray**, University of Massachusetts, Amherst
  "Feature-Weighted Ensembles for Probabilistic Time-Series Forecasts"
  Evan L. Ray, Nicholas G. Reich

  Accurate and reliable predictions of infectious disease incidence are important for public health decision makers planning resource allocation and interventions designed to prevent or reduce disease transmission. Ensemble prediction methods, which combine predictions from multiple "component" models, have recorded superior performance in a variety of tasks from

weather prediction to product recommendation; however, applications of ensemble methods in the context of predicting infectious disease have been limited. We considered a range of ensemble methods that each form a predictive density for a target of interest as a weighted sum of the predictive densities from several component models. In the simplest case, equal weight is assigned to each component model; in the most complex case, the weights vary with multiple observed features such as recent observations of disease incidence and the time of the year when predictions are made. We applied these methods to predict measures of influenza season timing and severity in the United States, both at the national and regional levels, using three component models. We trained the models on retrospective predictions from 14 seasons (1997/1998 - 2010/2011) and evaluated each model's prospective, out-of-sample performance in the five subsequent influenza seasons. In this test phase, the ensemble methods showed overall performance that was similar to the best of the component models, but offered more consistent performance across seasons than the component models. Ensemble methods offer the potential to deliver more reliable infectious disease predictions to public health decision makers.

- **Daeyoung Kim**, University of Massachusetts-Amherst
  "Confidence Distribution Sampling and Its Application"
  Daeyoung Kim

  Inference functions, such as the likelihood, are widely used throughout statistics. They have the virtue of providing methods for point estimation, set estimation, and hypothesis testing. Researchers often perform inferences based on the asymptotic theory. But, the asymptotic-based inferences may not be reliable if the amount of available information is not large relative to the number of parameters. To address this long-standing problem, we have developed the methodologies to assess the adequacy of using the asymptotic theory for finite-sample inference. They are based on confidence distribution sampling and volumetric error analysis.

- **Patrick Flaherty**, University of Massachusetts-Amherst
  "a Deterministic Global Optimization Method for Variational Inference"
  Hachem Saddiki, Andrew C. Trapp, Patrick Flaherty

  Variational inference methods for latent variable statistical models have gained popularity because they are relatively fast, can handle large data sets, and have deterministic convergence guarantees. However, in practice it is unclear whether the fixed point identified by the variational inference algorithm is a local or a global optimum. Here, we propose a method for constructing iterative optimization algorithms for variational inference problems that are guaranteed to converge to the ϵ-global variational lower bound on the log-likelihood. We derive inference algorithms for two variational approximations to a standard Bayesian Gaussian mixture model (BGMM). We present a minimal data set for empirically testing convergence and show that a variational inference algorithm frequently converges to a local optimum while our algorithm always converges to the globally optimal variational lower bound. We characterize the loss incurred by choosing a non-optimal variational approximation distribution suggesting that selection of the approximating variational distribution deserves as much attention as the selection of the original statistical model for a given data set.

- **Matthias Steinruecken**, University of Massachusetts-Amherst
  "Unraveling the Demographic History of Modern Humans using Full-Genome Sequencing Data"
  Matthias Steinruecken

Contemporary and ancient demographic structure in human populations has shaped the genomic variation observed in modern humans, and severely affected the distribution of functional and disease related genetic variation. Using next-generation sequencing technologies, researchers gather increasing amounts of genomic sequencing data for large samples in many different human population groups. These datasets present unprecedented opportunities to study genomic variation in complex demographic scenarios, and this area has received a lot of attention in recent years.

In this talk, I will present a method for the inference of demographic histories from full-genome sequencing data of multiple individuals developed by me and my collaborators. I will apply this method to a genomic dataset of Native American individuals to unravel the ancient demographic events underlying the peopling of the Americas. Morevoer, I will discuss a novel method for demographic inference that has the potential to improve inference especially in the recent past, which is of particular importance in the context of complex genetic diseases in humans.

9. Social Networks and Causal Inference

- **Daniel Sussman**, Boston University
  "Optimal Unbiased Estimation of Causal Effects under Network Interference"
  Daniel Sussman, Edo Airoldi

  From a causal inference perspective, the typical assumption of no interference becomes untenable in experiments in a social context. In many instances, however, the patterns of interference may be informed by the observation of network connections among the units of analysis. We develop elements of optimal estimation theory for causal effects by leveraging an observed network. Considering the class of linear unbiased estimators of the average direct treatment effect under various exclusion restrictions for the potential outcomes, we offer analytical insights on the weights that lead to minimum integrated variance estimators. These estimators offer superior performance to previously proposed estimators and we seek to develop complementary variance estimators for these estimates based on similar principles.

- **Alexander Volfovsky**, Duke University
  "Causal Inference in the Presence of Networks: Randomization and Observation"
  Alexander Volfovsky

  Much of classical causal analysis relies on notions of independence. However, modern datasets on disease prevalence, social development, online advertising and business transactions come equipped with information on a network that links the units together, rendering these notions implausible. When designing randomized experiments, scientists must control for network interference and homophily in order to guarantee the theoretical properties of their estimators. Studying the direct treatment effect in networks, we describe a new class of randomizations that can guarantee unbiasedness and control the variance of the estimator. In situations where an experiment cannot be performed, causal analysis requires the use of matching techniques in order to protect against bias due to a lack of balance between treated and control units. We provide examples of the complications that arise when information about the network is disregarded and develop a matching technique that extends classical propensity scores to the realm of networks.

- **Dean Eckles**, Mit
  "Estimating Peer Effects in Networks with Peer Encouragement Designs"
  Dean Eckles, Rene Kizilcec, Eytan Bakshy

  Peer effects, in which the behavior of an individual is affected by the behavior of their peers, are central to social science. Because peer effects are often confounded with homophily and common external causes, recent work has used randomized experiments to estimate effects of specific peer behaviors. These experiments have often relied on the experimenter being able to randomly modulate mechanisms by which peer behavior is transmitted to a focal individual. We describe experimental designs that instead randomly assign individuals peers to encouragements to behaviors that directly affect those individuals. We illustrate this method with a large peer encouragement design on Facebook for estimating the effects of receiving feedback from peers on posts shared by focal individuals. We find evidence for substantial effects of receiving marginal feedback on multiple behaviors, including giving feedback to others and continued posting. These findings provide experimental evidence for the role of behaviors directed at specific individuals in the adoption and continued use of communication technologies. In comparison, observational estimates differ substantially, both underestimating and overestimating effects, suggesting that researchers and policy makers should be cautious in relying on them.

- **Hyunseung Kang**, University of Wisconsin Madison
  "Peer Encouragement Designs in Causal Inference with Partial Interference and Identification of Local Average Network Effects"
  Hyunseung Kang, Guido Imbens

  In non-network settings, encouragement designs have been widely used to analyze causal effects of a treatment, policy, or intervention on an outcome of interest when randomizing the treatment was considered impractical or when compliance to treatment cannot be perfectly enforced. Unfortunately, such questions related to treatment compliance have received less attention in network settings and the most well-studied experimental design in networks, the two-stage randomization design, requires perfect compliance with treatment. The paper proposes a new experimental design called peer encouragement design to study network treatment effects when enforcing treatment randomization is not feasible. The key idea in peer encouragement design is the idea of personalized encouragement, which allows point-identification of familiar estimands in the encouragement design literature. The paper also defines new causal estimands, local average network effects, that can be identified under the new design and analyzes the effect of non-compliance behavior in randomized experiments on networks.

  This is joint work with Guido Imbens (Stanford)

10. Statistical Innovations in Genomics

- **Hongkai Ji**, Johns Hopkins University
  "Single-Cell Rna-Seq Analysis by Spanning Trees"
  Zhicheng Ji, Hongkai Ji

  When analyzing single-cell RNA-seq data, constructing a pseudo-temporal path to order cells based on the gradual transition of their transcriptomes is a useful way to study gene

expression dynamics in a heterogeneous cell population. We present TSCAN, a new tool for single cell pseudo-time analysis. TSCAN uses a cluster-based minimum spanning tree (MST) approach to order cells. Cells are first grouped into clusters and an MST is then constructed to connect cluster centers. Pseudo-time is obtained by projecting each cell onto the tree, and the ordered sequence of cells can be used to study dynamic changes of gene expression along the pseudo-time. Clustering cells before MST construction reduces the complexity of the tree space. This often leads to improved cell ordering. It also allows users to conveniently adjust the ordering based on prior knowledge. TSCAN has a graphical user interface (GUI) to support data visualization and user interaction. Furthermore, quantitative measures are developed to objectively evaluate and compare different pseudo-time reconstruction methods.

11. Recent Developments on High-Dimensional Statistics and Regularized Estimation

- **Ethan Fang**, Pennsylvania State University-Main Campus
  "Blessing of Massive Scale: Spatial Graphical Model Estimation with a Total Cardinality Constraint Approach"
  Ethan Fang, Han Liu, Mengdi Wang

  We consider the problem of estimating high dimensional spatial graphical models with a total cardinality constraint. Though this problem is highly nonconvex, we show that its primal-dual gap diminishes linearly with the dimensionality and provide a convex geometry justification of this "blessing of massive scale" phenomenon. Motivated by this result, we propose an efficient algorithm to solve the dual problem (which is concave) and prove that the solution achieves optimal statistical properties. Extensive numerical results are also provided.

- **Cheng Yong Tang**, Temple University
  "Sufficient Dimension Reduction with Missing Data"
  Yuexiao Dong, Cheng Yong Tang, Qi Xia

  Inverse regressions constitute a class of sufficient dimension reduction methods targeting at estimating the central space by regression-type approaches implemented inversely on the predictors and the responses. The most representative approach in this family is the seminal Sliced Inverse Regression (SIR) approach proposed by Li (1991). In this study, we first show that missing responses generally affect the validity of the inverse regressions under the scheme of the so-called missing at random, in the sense that the resulting estimations for the central space can be biased if data with missing responses are simply ignored. We then propose two simple and effective adjustments for missing responses that guarantees the validity of the inverse regressions. The proposed methods share the essence and simplicity of the inverse regressions. We demonstrate the performance of the proposed inverse regressions for dealing with missing responses by numerical and theoretical analyses.

- **Sahand Negahban**, Yale University
  "On Approximation Guarantees for Greedy Low Rank Optimization"
  Rajiv Khanna, Ethan R. Elenberg, Alexandros G. Dimakis, and Sahand Negahban

  We provide new approximation guarantees for greedy low rank matrix estimation under standard assumptions of restricted strong convexity and smoothness. Our novel analysis also uncovers previously unknown connections between the low rank estimation and combinatorial optimization, so much so that our bounds are reminiscent of corresponding approximation

bounds in submodular maximization. Additionally, we also provide statistical recovery guarantees.

- **Ting Zhang**, Boston University
  "a Thresholding-Based Prewhitened Long-Run Variance Estimator and Its Dependence-Oracle Property"
  Ting Zhang

  Statistical inference of time series data routinely relies on the estimation of long-run variances, defined as the sum of autocovariances of all orders. The current paper considers a new class of long-run variance estimators, which first soaks up the dependence by a decision-based prewhitening filter, then regularizes autocorrelations of the resulting residual process by thresholding, and finally recolors back to obtain an estimator of the original process. Under mild regularity conditions, we prove that the proposed estimator (i) consistently estimates the long-run variance; (ii) achieves the parametric convergence rate when the underlying process has a sparse dependence structure as in finite-order moving average models; and (iii) enjoys the dependence-oracle property in the sense that it will automatically reduce to the sample variance if the data are actually independent. Monte Carlo simulations are conducted to examine its finite-sample performance and make comparisons with existing estimators.

12. Subgroup Analysis

- **Yanxun Xu**, Johns Hopkins University
  "a Nonparametric Bayesian Basket Trial Design"
  Yanxun Xu, Peter Mueller, Apostolia Tsimberidou, Donald Berry

  Targeted therapies on the basis of genomic aberrations analysis of the tumor have shown promising results in cancer prognosis and treatment. Regardless of tumor type, trials that match patients to targeted therapies for their particular genomic aberrations have become a mainstream direction of therapeutic management of patients with cancer. Therefore, finding the subpopulation of patients who can most benefit from an aberration-specific targeted therapy across multiple cancer types is important. We propose an adaptive Bayesian clinical trial design for patient allocation and subpopulation identification. We start with a decision theoretic approach, including a utility function and a probability model across all possible subpopulation models. The main features of the proposed design and population finding methods are that we allow for variable sets of covariates to be recorded by different patients, adjust for missing data, allow high order interactions of covariates, and the adaptive allocation of each patient to treatment arms using the posterior predictive probability of which arm is best for each patient. The new method is demonstrated via extensive simulation studies.

- **Jared Huling**, University of Wisconsin-Madison
  "Heterogeneity of Intervention Effects and Subgroup Identification Based on Longitudinal Outcomes"
  Jared Huling, Menggang Yu, Maureen Smith

  With the growing health costs in the United States, the need to improve the efficiency and efficacy of care has become increasingly urgent. There has been great interest in developing interventions to effectively coordinate the typically fragmented care of patients with many comorbidities. Evaluation of such interventions is often challenging given their long-term

nature and their differential effectiveness among different patients. Given this and the resource intensiveness of care coordination interventions, there is significant interest in identifying which patients may benefit the most from care coordination. Identification of patients which benefit from a particular intervention can be accomplished by modeling covariates which modify the intervention effect. In this work we extend the interaction modeling framework of Tian, et al. (2014) and Chen, et al. (2017) to handle long-term interventions whose effects are expected to change smoothly over time. We allow interaction effects to vary over time and encourage these effects to be more similar over time by utilizing a fused lasso penalty. Our approach allows for flexibility in modeling temporal effects while also borrowing strength in estimating these effects over time. We use our approach to identify a subgroup of patients who benefit from a complex case management intervention in a large hospital system.

- **Wai-Ki Yip**, Foundation Medicine, Inc.
  "Stepp Analysis for Continuous, Binary, and Count Outcomes and Other Recent Stepp Development"
  Wai-Ki Yip, Marco Bonetti, Ann Lazar, William Barcella, Victoria Xin Wang, Chip Cole, Rich Gelber

  The Subpopulation Treatment Effect Pattern Plot is a visual and statistical technique to explore patterns of treatment effects across values of a continuously measured covariate such as a biomarker measurement. Originally developed specifically for investigation of survival outcomes, it has been extended to continuous, binary and count outcomes. This talk will focus on the development of this extension what are the outcomes, the permutation statistics and the Type 1 error, the power of the test, comparison with other methods, and the software. Then, a motivating example of how it is applied to analyze data from the Aspirin/Folate Polyp Prevention Study will be presented. A quick summary of recent research development in STEPP will be presented at the end.

Abstracts of Posters

Posters

- **Suzanne Thornton**, Rutgers, The State University of New Jersey
  "Approximate confidence distribution computing: An effective likelihood-free method with statistical guarantees"
  Suzanne Thornton, Min-ge Xie

  Approximate Bayesian computing (ABC) is a likelihood-free method that has grown increasingly popular since early applications in population genetics. However, the theoretical justification for inference based on this method has yet to be fully developed especially pertaining to the use of non-sufficient summary statistics. We introduce a more general computational technique, approximate confidence distribution computing (ACC) to overcome two defects of the ABC method, namely, lack of theory supporting the use of non-sufficient summary statistics and lack of guardian for the selection of prior. Specifically, we establish frequentist coverage properties for the outcome of the ACC method by using the theory of confidence distributions, and thus inference based on ACC is justified (even if reliant upon a non-sufficient summary statistic). Furthermore, the ACC method is very broadly applicable; in fact, the ABC algorithm can be viewed as a special case of an ACC method without damaging the integrity of ACC based inference. We supplement the theory with simulation studies and an epidemiological application to illustrate the benefits of the ACC method. It is also demonstrated that a well-tended ACC algorithm can greatly increase its computing efficiency over a typical ABC algorithm.

- **Shaoyang Ning**, Harvard University
  "A Nonparametric Bayesian Approach to Copula Estimation"
  Shaoyang Ning, Neil Shephard

  We propose a novel Dirichlet-based Pólya tree (D-P tree) prior on the copula and based on the D-P tree prior, a nonparametric Bayesian inference procedure. Through theoretical analysis and simulations, we are able to show that the flexibility of the D-P tree prior ensures its consistency in copula estimation, thus able to detect more subtle and complex copula structures than earlier nonparametric Bayesian models, such as a Gaussian copula mixture. Further, the continuity of the imposed D-P tree prior leads to a more favorable smoothing effect in copula estimation over classic frequentist methods, especially with small sets of observations. We also apply our method to the copula prediction between the S&P 500 index and the IBM stock prices during the 2007-08 financial crisis, finding that D-P tree-based methods enjoy strong robustness and flexibility over classic methods under such irregular market behaviors.

- **Xinran Li**, Harvard University
  "Asymptotic Theory of Rerandomization in Treatment-Control Experiments"
  Xinran Li, Peng Ding, Donald B. Rubin

  Although complete randomization ensures covariate balance on average, the chance for observing significant differences between treatment and control covariate distributions increases with many covariates. Rerandomization discards randomizations that do not satisfy a predetermined covariate balance criterion, generally resulting in better covariate balance and more precise estimates of causal effects. Previous theory has derived finite sample theory for

rerandomization under the assumptions of equal treatment group sizes, Gaussian covariate and outcome distributions, or additive causal effects, but not for the general sampling distribution of the difference-in-means estimator for the average causal effect. To supplement existing results, we develop asymptotic theory for rerandomization without these assumptions, which reveals a non-Gaussian asymptotic distribution for this estimator, specifically a linear combination of a Gaussian random variable and a truncated Gaussian random variable. This distribution follows because rerandomization affects only the projection of potential outcomes onto the covariate space but does not affect the corresponding orthogonal residuals. We also demonstrate that, compared to complete randomization, rerandomization reduces the asymptotic sampling variances and quantile ranges of the difference-in-means estimator. Moreover, our work allows the construction of accurate large-sample confidence intervals for the average causal effect, thereby revealing further advantages of rerandomization over complete randomization.

- **Jean Pouget-Abadie**, Harvard University
  "Randomizing over randomized experiments to test for network interference"
  Jean Pouget-Abadie, Martin Saveski, Guillaume Saint-Jacques, Weitao Duan, Ya Xu, Souvik Ghosh, Edoardo Maria Airoldi

  We propose an experimental design for testing whether the stable unit value assumption holds, by comparing two different estimates of the total treatment effect obtained through two different assignment strategies: a completely randomized assignment and a cluster-based randomized assignment. We provide a methodology for obtaining these two estimates simultaneously and provide theoretical guarantees for rejecting the null hypothesis that the stable unit value assumption holds without specifying a model of interference. We provide a discussion on how to apply our methodology to large internet experimentation platforms. Finally, we illustrate the proposed multilevel design to a live experiment on the LinkedIn platform.

- **Zach Branson**, Harvard University
  "A Nonparametric Bayesian Methodology for Analyzing Regression Discontinuity Designs"
  Zach Branson, Maxime Rischard, Luke Bornn, and Luke Miratrix

  Regression discontinuity designs (RDDs) are natural experiments where treatment assignment is determined by a covariate value (or "running variable") being above or below a predetermined threshold. Because the treatment effect will be confounded by the running variable, RDD analyses focus on the local average treatment effect (LATE) at the threshold, where treated and control units are most similar in terms of the running variable. The most popular methodology for estimating the LATE in an RDD is local linear regression (LLR), which is a weighted linear regression that places larger weight on units closer to the threshold. While LLR exhibits promising bias-reducing properties, LLR tends to yield confidence intervals that undercover, in part because LLR assumes the weighting scheme is fixed, when really there is uncertainty in the weighting scheme choice. We propose an alternative nonparametric methodology utilizing Gaussian process regression that, unlike LLR, (1) does not require specifying a functional form for the expected treatment and control response, and (2) automatically incorporates the uncertainty in how units are weighted when estimating the treatment effect. We prove our methodology is consistent for the LATE, and we replicate previous simulation studies in the literature to show that our method exhibits better coverage and mean square error properties than current methodologies.

- **Elizabeth Upton**, Boston University
  "Bayesian Network Regularized Regression for Modeling Urban Crime Occurrences"
  Elizabeth Upton, Luis Carvalho

  We consider the problem of statistical inference and prediction for processes defined on networks. We assume that the network is known and measures similarity, and our goal is to learn about an attribute associated with its vertices. Classical regression methods are not immediately applicable to this setting, as we would like our model to incorporate information from both network structure and pertinent covariates. Our proposed model consists of a generalized linear model with vertex indexed predictors and a basis expansion of their coefficients, allowing the coefficients to vary over the network. We employ a regularization procedure, cast as a prior distribution on the regression coefficients under a Bayesian setup, so that the predicted responses vary smoothly according to the topology of the network. We first motivate the need for this model by examining occurrences of residential burglary in Boston, Massachusetts. Noting that crime rates are not spatially homogeneous, and that the rates appear to vary sharply across regions or hot zones in the city, we construct a hierarchical model that addresses these issues and gives insight into spatial patterns of crime occurrences. Furthermore, we examine an efficient expectation-maximization fitting algorithm and provide computationally-friendly methods for eliciting hyper-prior parameters. We demonstrate the performance of the proposed model in a simulation study and a case study in Boston.

- **Qin Lu**, University of Connecticut
  "The Multidimensional Cramer-Rao-Leibniz Lower Bound for Vector-Meaurement-based Likelihood Functions with Parameter-Dependent Support"
  Qin Lu, Yaakov Bar-Shalom, Peter Willett, Francesco Palmieri, Fred Daum

  One regularity condition for the classical Cramer-Rao lower bound (CRLB) of an unbiased estimator to hold — that the support of the likelihood function (LF) should be independent of the parameter to be estimated — has recently been relaxed to the case of parameter-dependent support as long as the LF is continuous at the boundary of its support. For the case where the LF is not continuous on the boundary of its support, a new modified CRLB — designated the Cramer-Rao-Leibniz lower bound (CRLLB) as it relies on the Leibniz integral rule — has also been presented for the scalar parameter case. The present work derives the multidimensional CRLLB for the case of LF based on vector measurements with parameter-dependent support by applying the general Leibniz integral rule to complete the framework of the CRLLB. Some illustrative examples have been provided to demonstrate the evaluation of the CRLLB.

- **Tom Chen**, Harvard University
  "A stochastic second-order generalized estimating equations approach for estimating intraclass correlation in the presence of informative missing data"
  Tom Chen, Eric J. Tchetgen Tchetgen, Rui Wang

  Design and analysis of cluster randomized trials must take into account correlation among outcomes from the same clusters. When applying standard generalized estimating equations (GEE), the first-order (e.g. treatment) effects can be estimated consistently even with a misspecified correlation structure. In settings for which the correlation is of interest, one could estimate this quantity via second-order generalized estimating equations (GEE2). We build upon GEE2 in the setting of missing data, for which we incorporate a "second-order" inverse-probability weighting (IPW) scheme and "second-order" double robustness (DR) equations

that guard against model misspecification. We highlight the need to model correlation among missingness indicators in such settings. In addition, the computational difficulties in solving these second-order equations have motivated our development of stochastic algorithms for solving GEE2s, which alleviates the reliance on starting points and provides substantially faster convergence and a higher convergence rate than deterministic root-solving methods.

- **Jessica Hoag**, University of Connecticut
  "Hemoglobinopathies and adverse cancer-related outcomes: A multi-technique approach for analyzing tumor registry data linked to Medicare claims"
  Jessica Hoag, Biree Andemariam, Xiaoyan Wang, David Gregorio, Helen Swede

  Racial disparities in cancer outcomes persist despite recent improvements in mortality, prompting investigations into the prognostic interplay of biological, individual, and social factors. Preclinical and case report evidence have shown that malformed red blood cells present in individuals with inherited hemoglobin variants such as sickle cell trait can interact with the tumor microenvironment to induce treatment failure and systemic adverse events. Hemoglobinopathies are disproportionately prevalent among African American/Blacks (AA/B) compared to non-Hispanic whites (NHW), but the distinct and synergistic effects of hemoglobin variants with treatment completion and adverse events on cancer survival is unknown.

  Given this lack of background understanding, multiple statistical approaches were employed to quantify the contribution of hemoglobinopathies to black-white differences in cancer-related outcomes in a large observational study cohort.

  We identified 162,357 older breast (n=75,633) and prostate (n=86,904) cancer patients diagnosed 2007-2013 using the SEER-Medicare linked database. AA/B and NHW patients were grouped by hemoglobinopathy status (AA/B+, AA/B-, NHW-) and three-way propensity score weighting using generalized boosted models (GBM) was performed to control for imbalances in demographic and clinicopathological features across study groups. The relative risk (RR) of treatment failure and occurrence of one or more adverse event was modeled using a modified Poisson regression approach with robust error variance, and interactions between treatment completion and adverse events by hemoglobinopathy status were evaluated for their relative contributions to all-cause, cancer-specific, and competing risks survival.

  After propensity score weighting, no significant association was observed in treatment completion status between AA/B+ and AA/B- or NHW-. Among treated patients, however, AA/B+ status conferred increased RR of experiencing one or more adverse event compared to either AA/B- (RR: 1.15, 95

- **Yeongjin Gwon**, University of Connecticut
  "Network Meta-Regression for Ordinal Outcomes: Applications in Comparing Crohns Disease Treatments"
  Yeongjin Gwon, May Mo, Ming-Hui Chen, Juan Li, H. Xia Amy, Joseph Ibrahim

  Crohns Disease is a life-long condition associated with recurrent relapses characterized by abdominal pain, weight loss, anemia, and persistent diarrhea. In the U.S., there are approximately 780,000 Crohns disease patients and 33,000 new cases are added each year. In this paper, we propose a new network meta-regression approach for modeling ordinal outcomes in order to assess the ecacy of treatments for Crohns disease. Specically, we develop regression models based on aggregate trial-level covariates for the underlying cut-o points of the ordinal outcomes as well as for the variances of the random eects to capture heterogeneity

across trials. Our proposed models are particularly useful for indirect comparisons of multiple treatments that have not been compared head-to-head within the network meta-analysis framework. Moreover, we introduce Pearson residuals to detect outlying trials and construct an invariant test statistic to evaluate goodness-of-t in the setting of ordinal outcome metadata. A detailed case study demonstrating the usefulness of the proposed methodology is carried out using aggregate ordinal outcome data from 16 clinical trials for treating Crohns disease.

- **Yujing Jiang**, University of Connecticut
  "Fingerprinting Changes in Climate Extremes with Joint Modeling of Observations and Climate Model Simulation"
  Yujing Jiang, Jun Yan, Xuebin Zhang

  Detection and attribution (D&A) analysis for climate extremes plays an important role in understanding the human influence on the observed change in climate extremes. Recent developed methodologies for D&A analysis use signal estimated from climate model simulation under external forcing as covariate in the model of observed extremes and carry out statistical analysis on the coefficient of the signal. The estimated signal contains statistical error, however, which may yield bias in the following analysis. In this study, we propose a method which combines the two stages of signal estimation and D&A analyis, and estimate the signal jointly from both the simulated and the observed extremes. We show that this method can reduce the bias effectively in the estimation compared to the previous method using a simulation study.

- **Phyllis Wan**, Columbia University in the City of New York
  "Threshold Selection for Multivariate Heavy-Tailed Data"
  Phyllis Wan, Richard A. Davis

  Regular variation is often used as the starting point for modeling multivariate heavy-tailed data. A random vector is regularly varying if and only if its radial part R is regularly varying and is asymptotically independent of the angular part Θ as R goes to infinity. The conditional limiting distribution of Θ given R is large characterizes the tail dependence of the random vector and hence its estimation is the primary goal of applications. A typical strategy is to look at the angular components of the data for which the radial parts exceed some threshold. While a large class of methods has been proposed to model the angular distribution from these exceedances, the choice of threshold has been scarcely discussed in the literature. In this paper, we describe a procedure for choosing the threshold by formally testing the independence of R and Θ using a measure of dependence called distance covariance. We generalize the limit theorem for distance covariance to our unique setting and propose an algorithm which automatically selects the threshold for R. This algorithm incorporates a subsampling scheme, which avoids the heavy computation in the calculation of the distance covariance, a typical disadvantage for this measure. The performance of our method is illustrated on both simulated and real data.

- **Kendra Plourde**, Boston University
  "Differences in Estimation between the Longitudinal Model and the Longitudinal portion of the Joint Model"
  Kendra Plourde, Yorghos Tripodis

  We investigate the effect of a joint survival and longitudinal models on the precision and accuracy of the longitudinal estimates. Mixed effects analysis has allowed investigators to in-

corporate more information in their models by allowing subjects to have repeated measures. More recently, joint models consisting of a cox proportional hazards model and a longitudinal mixed effects model have been proposed allowing investigators to additionally incorporate time-to-event data. By incorporating more information, we expect the estimates of the longitudinal portion of the joint model to be less biased and more precise on average. Extensive research has been done to show the improvement in estimation of the hazard function using joint models, but not much research has been done to investigate the differences in estimation of the longitudinal model. In this study, we compared the longitudinal model with the longitudinal portion of the joint model in terms of coverage, bias, and precision using the same simulation structure used previously (Mayeda, 2015). Our results showed that although the estimate of the longitudinal portion of the joint model was on average more precise, it had a higher root mean square error and was more susceptible to survival bias and type I error compared to the longitudinal model alone.

- **Dongah Kim**, University of Massachusetts-Amherst
  "Multivariate association in Respondent-Driven Sampling data"
  Dongah Kim, Krista J.Gile, Pedro Mateu-Gelabert, Honoria Guarino

  Respondent-Driven Sampling (Heckathorn 1997) is a sampling method designed to collect data for hard-to-reach populations; injected drug users, sex workers, and man who have sex with man. Beginning with a convenience sample, the sample recruits other participants using small number of uniquely-identified coupons to distribute among his/her social network. Coupon recipients can accept or reject participation of the survey study, and he/she also get small number of coupon to recruit other participants. Using these process, survey team can reach a desire sample size of the target population. This method is very effective to collect a data for hard-to-reach populations. However, valid statistical inference for these kinds of data relies on many strong assumptions. Most of all, statistical tests for between pairs of variables has strong limitation. In standard survey samples, we can assume the two pairs of variables from each individual are independent. In RDS condition, however, this assumption does not be satisfied because of the sampling dependence between individuals. Therefore, we propose to design methods to non-parametrically estimate the null distributions of standard test statistics in the presence of sampling dependence, allowing for more valid statistical testing.

- **Gregory Vaughan**, University of Connecticut
  "Efficient Interaction Selection via Stagewise Generalized Estimating Equations"
  Gregory Vaughan, Robert Aseltine, Kun Chen, Jun Yan

  Stagewise estimation is a slow-brewing approach for model building that has recently experienced a revival due to its computational efficiency, its flexibility in handling complex data structure, and its intrinsic connections with penalized estimation. Built upon generalized estimating equations, we propose general stagewise estimation approaches for variable and interaction selection in non-Gaussian/non-linear models with clustered data. As it is often required that main effect terms be included when an interaction term is part of a model, the key is to perform variable selection that maintains the variable hierarchy. We develop two techniques to address this challenge. The first is a hierarchical lasso stagewise estimating equations (hlSEE) approach, which is shown to directly correspond to the hierarchical lasso penalized regression. The second is an interaction stagewise estimating equations (iSEE) approach, which enforces the variable hierarchy by conforming the selection to a properly growing active set in each stagewise estimation step. Simulation studies are presented to

show the efficacy and superior computational efficiency of the proposed approaches. We apply the proposed approaches to study the association between the suicide-related hospitalization rates of the 15–19 age group and the characteristics of the school districts in the State of Connecticut.

- **Daoyuan Shi**, University of Connecticut
  "New Partition Based Measures for Data Compatibility and Information Gain"
  Daoyuan Shi, Lynn Kuo, Ming-Hui Chen

  It is of great practical importance to compare and combine data from different studies in order to carry out appropriate and more powerful statistical inference. In this paper, to quantify the compatibility of two data sets we first propose a partition based measure in terms of the corresponding posterior distributions of the parameters. We further propose an information gain to measure the information increase in combining two data sets. These measures are well calibrated. Efficient computational algorithms are developed for calculating these measures. We illustrate how these two measures are useful in combining historical data to current data with a benchmark toxicology example.

- **Suzanne Thornton**, Rutgers University-New Brunswick
  "Approximate confidence distribution computing: An effective likelihood-free method with statistical guarantees"
  Suzanne Thornton, Min-ge Xie

  Approximate Bayesian computing (ABC) is a likelihood-free method that has grown increasingly popular since early applications in population genetics. However, the theoretical justification for inference based on this method has yet to be fully developed especially pertaining to the use of non-sufficient summary statistics. We introduce a more general computational technique, approximate confidence distribution computing (ACC) to overcome two defects of the ABC method, namely, lack of theory supporting the use of non-sufficient summary statistics and lack of guardian for the selection of prior. Specifically, we establish frequentist coverage properties for the outcome of the ACC method by using the theory of confidence distributions, and thus inference based on ACC is justified (even if reliant upon a non-sufficient summary statistic). Furthermore, the ACC method is very broadly applicable; in fact, the ABC algorithm can be vi ewed as a special case of an ACC method without damaging the integrity of ACC based inference. We supplement the theory with simulation studies and an epidemiological application to illustrate the benefits of the ACC method. It is also demonstrated that a well-tended ACC algorithm can greatly increase its computing efficiency over a typical ABC algorithm.

- **Qiongshi Lu**, Yale University
  "A powerful approach to estimating annotation-stratified genetic covariance using GWAS summary statistics"
  Qiongshi Lu, Boyang Li, Derek Ou, Margret Erlendsdottir, Ryan Powles, Tony Jiang, Yiming Hu, David Chang, Chentian Jin, Wei Dai, Qidu He, Zefeng Liu, Shubhabrata Mukherjee, Paul Crane, Hongyu Zhao

  Despite the success of large-scale genome-wide association studies (GWASs) on complex traits, our understanding of their genetic architecture is far from complete. Jointly modeling multiple traits genetic profiles has provided insights into the shared genetic basis of many complex traits. However, large-scale inference sets a high bar for both statistical power and biological interpretability. Here we introduce a principled framework to estimate annotation-stratified

genetic covariance between traits using GWAS summary statistics. Through theoretical and numerical analyses we demonstrate that our method provides accurate covariance estimates, thus enabling researchers to dissect both the shared and distinct genetic architecture across traits to better understand their etiologies. Among 50 complex traits with publicly accessible GWAS summary statistics (Ntotal 4.5 million), we identified more than 170 pairs with statistically significant genetic covariance. In particular, we found strong genetic covariance between late-onset Alzheimers disease (LOAD) and amyotrophic lateral sclerosis (ALS), two major neurodegenerative diseases, in single-nucleotide polymorphisms (SNPs) with high minor allele frequencies and in SNPs located in the predicted functional genome. Joint analysis of LOAD, ALS, and other traits highlights LOADs correlation with cognitive traits and hints at an autoimmune component for ALS.

- **David Cheng**, Harvard T.H. Chan School of Public Health
  "Efficient and Robust Semi-Supervised Estimation of Average Treatment Effects in Electronic Medical Records Data"
  David Cheng, Ashwin Ananthakrishnan, Tianxi Cai

  There is strong interest in conducting comparative effectiveness research (CER) in electronic medical records (EMR) data to evaluate treatment strategies among real-world patients. A primary challenge of working with EMR data is the lack of direct observation on a pre-specified true outcome, prompting the need for phenotyping algorithms that impute the outcome given available data. It is often unclear whether such imputations are adequate when used to estimate the treatment effect. We frame the problem of estimating average treatment effects (ATE) in a semi-supervised learning setting, where we suppose a small set of observations labeled with the true outcome and a large set of unlabeled observations are available. We develop an approach for imputing the outcome and an estimator for the ATE that such that the treatment effect estimator is robust to mis-specification of the imputation model. As a result, information from surrogate variables that predict the outcome in the unlabeled data can safely be leveraged to improve the efficiency in estimating the ATE. The estimator is also doubly-robust in that it will be consistent under correct specification of either an initial propensity score model or a baseline outcome model. It is more efficient than complete-case estimators that neglect the unlabeled data and related missing data and causal inference estimators we adapt to this setting to make use of the unlabeled data. Simulations exhibit the efficiency and robustness benefits of the proposed estimator in finite samples. We illustrate the method in an EMR study to compare rates of treatment response to two anti-TNF therapies for the management of inflammatory bowel disease.

- **Wenjie Wang**, University of Connecticut
  "Extended Cox Model by ECM Algorithm for Uncertain Survival Records Due to Imperfect Data Integration"
  Wenjie Wang, Kun Chen, Jun Yan

  In the era of big data, there has been an increasing need in using data integrated from disparate sources to conduct statistical analysis. The potential benefits from data integration, however, may be compromised by the induced data uncertainty due to incomplete/imperfect linkage, causing potential bias in statistical inference. It is thus pivotal to take into account the uncertainty associated with data integration. Motivated by a suicide prevention study, we consider a survival analysis setup to handle uncertainty event records arising from data integration. Specifically, a survival dataset constructed from hospital discharge fails to capture the events of interest for all the subjects, and the missing events may be recovered from

a complete death record database that contains all the event records of a much larger population. Nonetheless, the original dataset can only be linked to the database by matching basic characteristics of subjects. As such, a censored subject from the original dataset could find multiple possible event times in the second database, which may or may not contain the true event time. We propose an extended the Cox regression approach, in which such uncertainty and mismeasurement of survival data are modeled probabilistically. The estimation procedure is derived in the spirit of expectation conditional maximization (ECM) algorithm and profile likelihood function. It takes regular the Cox model as a special case and reduces to the Cox model when there is not uncertainty in the data. The performance of the proposed method is evaluated through simulation studies. The proposed method outperforms the naive approaches under slight and severe censoring when the data matching leads to more true outcomes than noise. We show that the extend Cox model is practically attractive by applying it to the 2005–2012 suicide attempt data from the State of Connecticut, which suggests interesting and insightful results.

- **Michael C. Burkhart**, Brown University
  "The discriminative Kalman filter for nonlinear and non-Gaussian sequential Bayesian filtering"
  Michael C. Burkhart, David M. Brandman, Matthew T. Harrison

  The Kalman filter is used in a variety of applications for computing the posterior distribution of latent states in a state space model. The model requires a linear relationship between states and observations. Extensions to the Kalman filter have been proposed that incorporate linear approximations to nonlinear models such as the extended Kalman filter (EKF) and the unscented Kalman filter (UKF). However, we argue that in cases where the dimensionality of observed variables greatly exceeds the dimensionality of state variables, a model for $p(\text{state}|\text{observation})$ proves both easier to learn and more accurate for latent state estimation. We derive and validate what we call the discriminative Kalman filter (DKF): a closed-form discriminative version of Bayesian filtering that readily incorporates off-the-shelf discriminative learning techniques. We demonstrate how highly non-linear models for $p(\text{state}|\text{observation})$ can be specified. We validate on synthetic datasets. Finally, we discuss how the DKF has been successfully implemented for neural filtering in human volunteers in the BrainGate clinical trial.

- **Xinyu Chen**, Worcester Polytechnic Institute
  "Restricted Inference In Multiple Linear Regression"
  Xinyu Chen

  Regression analyses constitutes an important part of the statistical inference and has great applications in many areas. In some applications, we strongly believe that the regression function changes monotonically with some or all of the predictor variables in a region of interest. Deriving analyses under such constraints will be an enormous task. In our work, the restricted prediction interval for the mean of the regression function is constructed when two predictors are present. We use a modified likelihood ratio test(LRT) to construct confidence and prediction intervals.

- **Benedict Wong**, Harvard University
  "A Bayesian Approach to Correcting for Risk Factor Misclassification in the Partial Population Attributable Risk"
  Benedict Wong

Estimation of the population attributable risk (PAR) has become an important goal in public health research, because it describes the proportion of disease cases that could be prevented if an exposure were entirely eliminated from a target population as a result of some intervention. In epidemiological studies, categorical covariates are often misclassified. We present methods for obtaining point and interval estimates of the PAR in the presence of misclassification, using a Bayesian approach to estimate parameters of the logistic regression models for the disease and for the misclassification process, under two different study designs. We compare this method to a likelihood-based method in a simulation study, using estimates from data in the Health Professionals Follow-Up Study of risk factors for colorectal cancer.

- **Timothy Leonard**, University of Rhode Island
  "Predicting Authorship with Assortative Mixture of English Parts of Speech"
  Timothy Leonard

  Authorship attribution is a classification problem with two main objectives: 1) to accurately predict some characteristic of a piece of text (e.g. authorship), and 2) to provide a descriptive model of writing that contributes to our knowledge of language. This article presents an assortative mixture model of English parts of speech that accurately predicts authorship in a supervised learning environment. By measuring the tendency for same parts of speech to collocate, the model offers a detailed and unbiased glimpse into the stylometric features of grammar. Assortative mixture is a single coefficient that can be applied to each part of speech in a word graph to generate a small but inclusive feature set. Comprised of only a single estimator, the assortative mixture model is simple yet captures many fundamental language characteristics including what grammar types exhibit selective linking. As a network graph model, words are vertices and edges represent sequential words (i.e. word bigrams or adjacencies) that appear in a sample of writing. To calculate assortativity and generate a feature set, vertices have as an attribute a part of speech that can be compared to other vertices. Such graphs are not new to the literature, however, previous models ignore grammar or fail to represent all grammar types due to computational limitations or deliberate choice of the model. Research on word graphs sought to discover predictive features using network analysis but did not include the part of speech as an attribute of a vertex. These studies showed that other descriptive characteristics such as transitivity, density, degree assortativity, etc., do not stand alone as significant predictors in a feature set. By comparison, grammar assortativity alone is highly predictive of authorship. The statistical analysis of graphs aided with an accurate speech tagger empowers a more mathematically descriptive examination of grammar now that entire collections of writing can be tagged efficiently.

- **Jinxin Tao**, Worcester Polytechnic Institute
  "Comparison between confidence intervals of multiple linear regression models with and without restriction"
  Jinxin Tao Thelge Buddika Peiris

  Regression analysis is one of the most applied statistical techniques. The statistical inference of a linear regression model with a monotone constraint had been discussed in early analysis. A natural question arises when it comes to the difference between the cases of with and without the constraint. Although the comparison between confidence intervals of linear regression models with and without restriction for one predictor variable had been considered, this discussion for multiple regression is required. We discuss the comparison of the intervals between a multiple linear regression model with and without constraints.

- **Yishu Xue**, University of Connecticut
  "Tests and Diagnostics for Cox Proportional Hazards Model in the Online Updating Setting"
  Yishu Xue, Jun Yan and Elizabeth Schifano

  For big survival data, conventional estimation methods for Cox proportional hazards models may become infeasible with standard routines because of the limit of computer memory. With big stream data partitioned into blocks, an online updating approach with cumulative estimating equations and cumulatively updated estimating equations is developed with minimal storage requirement to estimate the Cox model parameters. The methodology naturally leads to an online updating test procedure for the proportional hazards assumption. For certain specific alternatives, window versions of the cumulative statistic help to detect change point sooner. The proposed test procedures is shown to be hold their sizes and have substantial power in simulation studies. The usefulness of the methods is illustrated with a real application to prostate cancer.

- **Indrani Mandal**, University of Rhode Island
  "Correlation analysis of multivariate Smartwatch data"
  Indrani Mandal, Debanjan Borthakur

  The advanced smartwatch has multiple functionalities that includes measurements of physiological parameters such as heart rate, galvanic skin resistance(GSR), temperature, acceleration, etc.Analysis of Multifaceted sensor data provides us with the scope of tracking physiological, behavioral and environmental information.This work aims to find the correlation between the various sensor modalities using standard multivariate time series analysis.The correlation analysis proposed in this paper can be instrumental in differentiating actual medical condition and an artifact anomaly.This analysis can bring new insight into the possible hidden relationship between the sensor modalities and hence can be useful in recognizing the neurological state of the user.

- **Katherine Abramski**, University of Rhode Island
  "A Network Based Analysis of the European Refugee Crisis"
  Katherine Abramski, Katio Dio, Natallia Katenka

  Europe is in the midst of the largest migration crisis the world has seen since WWII. Largely fueled by on-going conflicts in Africa and the Middle East, this mass migration is a multidimensional problem with numerous factors at play. To better understand this problem, we created a network representation of migration patterns between chosen countries in Europe, Africa, and the Middle East. Specifically, we observed the total number of refugees that migrated from one country to another in 2010 and 2015. One of our goals was to observe how migration patterns have changed over time and to understand what major events may have contributed to those changes. We also investigated various factors that could presumably influence migration patterns. We analyzed multiple descriptive measures of our network including degree and strength distribution, transitivity, assortativity, betweenness, graph partitioning, and node and edge attributes. We observed an overall increase in migration from 2010 to 2015, although we did observe a decrease in migration to and from some countries as well. For the most part, the changes we observed were consistent with recent world events. In the future, we plan to explore the use of dynamic networks, specifically, stochastic actor-based models for modelling the evolution of migration patterns over time. We also plan to explore the predictive capabilities of additional variables such as illegal border crossings, attitudes towards refugees, and legislation.

- **Gabriel De Pace**, University of Rhode Island
  "Applying CNNs to Human Facial Expressions for Emotion Recognition"
  Gabriel De Pace, Terry Ferguson, Indrani Mandal

  Automating emotion recognition has many valuable applications today, especially in light the proliferation of mobile devices capable of capturing and processing images. Health monitoring, mental health follow up and marketing could see benefits with a low-cost and fast implementation of emotion detection.

  We have used Caffe and OpenFace, open source CNN libraries and tools, to classify emotions using human facial expressions. We have used pre-trained deep learning models in order to establish benchmarks. We apply existing and newly trained networks to video data.

  Preliminarily, we have found that happiness is the easiest to classify. The distinction between surprise and fear is difficult to make. Disgust is the hardest to classify.

- **Anton Lobach**, University of Rhode Island, Dept. of Computer Science and Statistics
  "A Markov Switching Causal-Noncausal Autoregressive Model with Application to Economic Bubbles"
  Anton Lobach, Gavino Puggioni

  Speculative bubble phenomena may result in explosive trends followed by a sharp decline in financial time series. A bubble is formed when investors future profits expectations influence the present market value of securities. Mixed causal-noncausal autoregressive processes are able to better model such behavior in comparison to traditional ARIMA models. In this work we propose Markov switching mixed causal-noncausal autoregressive processes (MSMAR) to account for changes in regime at different times. Parameter estimation is conducted in a Bayesian framework via MCMC algorithms. The model is tested for performance with a simulation study and then applied to Bitcoin/USD exchange rate and US inflation data.

- **Isabel Nowinowski**, University of Rhode Island, Statistics
  "The Deadliest Days for Drivers: A Bayesian Nonparametric Analysis of Fatal Car Accidents"
  Isabel Nowinowski, Katie Abramski, Hilary Aroke, Kaitlin Dio, Eugene Quinn, Daniel M. Smith, Gavino Puggioni

  Highway road signs displaying witty messages promoting safe driving are not an uncommon sight on the morning commute in New England, especially around major holidays. As departments of transportation and public health and safety place more emphasis on safe driving, understanding factors that affect the occurrence of motor vehicle accidents is becoming increasingly relevant. The aim of this analysis was to investigate periodic day of week and day of year effects and the influence of certain holidays on counts of fatal accidents in the United States. Using data from the Fatality Analysis Reporting System (FARS), we examined patterns in the number of fatal car accidents from 1975 to 2015 using a nonparametric Gaussian process model with additive components. Special days included in the analysis were New Years Eve and Day, Valentines Day, Memorial Day, Fourth of July, Labor Day, Halloween, Thanksgiving, and Christmas. Our results demonstrate a day of week effect, a day of year effect, and an effect of certain special days. This approach can be used to inform the timing of interventions and allocation of resources aimed at reducing road traffic accidents at particular times of the week, year, and around holidays.

- **Kaitlin Dio**, University of Rhode Island
  "Exploring Feedback in an Introductory Biostatistics Course: A Repeated Measures Analysis"
  Kaitlin Dio, Natallia Katenka

  The generation of millennials is phasing out of undergraduate courses and the next generation is replacing them. Titled Generation Z these students were born into a world with technology where their phones contained the answers to nearly any question they could ask. In response to this changing landscape of learners and their new educational needs, we implemented additional feedback procedures (weekly quizzes with varying grade incentives, an introductory survey and two SATS-36 surveys) in an introductory biostatistics course at the University of Rhode Island. Our purpose was two-fold. First, explore the effect of grade incentives on weekly quizzes by building a piecewise linear mixed effect longitudinal model. Next, investigate the change in attitudes throughout the semester using multiple linear regression to control for the starting attitude of the students. Results from this analysis will be extended to a larger goal to model URI undergraduate students attitudes towards statistics and its relationship with course outcomes.

- **Joseph Langan**, University of Rhode Island
  "Fracking Activity and Earthquakes in Oklahoma, 2011-2015"
  Joseph Langan, Divana Boukari, Marjana Catanzaro, Ayako Miura, John Ragland, Michael Weir, Xin Zhou, Gavino Puggioni

  One of the most contentious and pressing environmental issues today facing oil-producing states in the central US is the interaction of fracking activity with earthquake frequency. Exploratory analysis of fracking and earthquake data collected in Oklahoma from 2011 to 2015 suggest that gas extraction is closely associated with seismically active fault lines. In this analysis, generalized linear models are used with two main objectives. The first object is to model the number of earthquakes as a function of monthly fracking wastewater injection volume and location; second, to investigate the potential link between earthquake magnitude and injection volumes, while controlling for location within the state. Inference is carried out within the Bayesian framework using MCMC algorithms. One of the main challenges of this analysis is harmonizing several data sources recorded at different spatial scales. The results of this work aim to provide a deeper understanding of how induced earthquakes are related to wastewater injection volumes and assist in improved spatial planning of fracking activity in order to strike a balance between Oklahoma's energy industry and the protection of its community and infrastructure.

- **Alicia S Chua**, Boston University
  "Incorporation of time series method for latent trajectory of longitudinal Mini-Mental State Examination in Alzheimers patients"
  Alicia S Chua, Yorghos Tripodis

  In Alzheimers patients, the Mini-Mental State Examination (MMSE) is a commonly used measure of general cognitive function. Longitudinal cognitive patient reported outcomes (CPRO) in Alzheimer patients are often highly nonlinear. This nonlinear feature of the data violates the major assumptions of commonly used statistical models such as linear mixed-effect (LME) models and generalized estimating equations (GEE) models. An alternative approach to this issue involves the incorporation of time series methods for the latent trajectory in multivariate time series of the MMSE. Data from 432 participants of the National Alzheimers Coordinating Center were used in this study. We examined the trajectories of participants

MMSE across study visits (up to 8 visits) via time series plots. Next, we examined the assumptions of a LME model for the data, and investigated alternative models that correctly explains MMSE in our cohort. Time series plots revealed violations of the linearity assumption of MMSE as an outcome over time, and the plots appeared to exhibit a certain trend and random effects across time. When we fit an LME model, we observed no significant association for MMSE over time (p=0.3884). However, when we fit an autoregressive integrated moving average (ARIMA) model, namely ARIMA (1,1,1), we observed a significant improvement in model fit in comparison to the LME model. The findings from this study may deepen our understanding on how to unbiasedly and efficiently model longitudinal CPRO data in Alzheimer patients and provide more insights in the pathway of disease at the initial pre-symptomatic stage of Alzheimers disease.

NESS 2017 Participants

Katherine Abramski
Abidemi Adeniji
Edoardo Airoldi
Yasuo Amemiya
Flavio De Angelis
Claudio Antonini
Brien Aronov
Laurence Aucella
Sudipto Banerjee
Sudeep Bapat
Haim Bar
Pierre c Bellec
Donald Bennett
Dilli Bhatta
Xuan Bi
Jacob Bien
Abhishek Bishoyi
Debanjan Borthakur
Divana Boukari
Emily Bowers
Zach Branson
Michael Burkhart
Erin Butts
Xizhen Cai
Joseph Cappelleri
Marcel Carcea
Marjana Catanzaro
Tom Chen
Xinyu Chen
Renjie Chen
Ming-Hui Chen
Kun Chen
David Cheng
Steven Chiou
Alicia Chua
Erin Conlon
Forrest Crawford
Debasmita Das
Victor Hugo Lachos Davila
Qiqi Deng
Dipak Dey
Kaitlin Dio
Dean Eckles
Birol Emir
Ethan Fang
Fei Fang
Patrick Flaherty
Paul Frydryk
Eugene Gallagher
Victoria Gamerman
Guojun Gan
Feng Gao
Beatriz Etchegaray Garcia
Miaomiao Ge
Lijiang Geng
Joseph Glaz
Gyuhyeong Goh
Yuwen Gu
Jiang Gui
Yeongjin Gwon
Aritra Halder
Ofer Harel
Jessica Hoag
Chaoran Hu
Jun Hu
Yuan Huang
Jared Huling
Hongkai Ji
Xiaoyu Jia
Yujing Jiang
Jieying Jiao
Shuaimin Kang
Sangwook Kang
Hyunseung Kang
Chihwa Kao
Vishesh Karwa
David Keller
Chanmin Kim
Dongah Kim
Daeyoung Kim
Siddhesh Kulkarni
Lynn Kuo
Nathan Lally
Matt Lamoureux
Joseph Langan
Liliya Lavitas
Kuang-Yao Lee
Timothy Leonard
Jerry Lewis
Steve Lewitzky
Xinran Li
Zhenghuan Li
Yujia Li
Mingjia Li
Yan Li
Xiaomeng Li
Eva Yujia Li
Xiaoran Li
Guannan Liang
Shuyi Liang
Kristin Lieber
Chaoyu Lin
Henry Linder
Kangyan Liu
Dungang Liu
Yang Liu
Xiao Liu
Ruiyin Liu
Ray Liu
Xiaokang Liu
Qingyang Liu
Greta Ljung
Anton Lobach
Rebecca Lough
Qin Lu
Qiongshi Lu
Chongliang Luo
Li Ma
Bani Mallick
Himel Mallick
Binod Manandhar
Indrani Mandal
Disheng Mao
Steven Martinez
Carolyn Mccormick
Paul Mclaughlin
Qian Meng
Ayako Miura
Shariq Mohammed
Oleksii Mostovyi
Jinjian Mu
Nitis Mukhopadhyay
Karthyek Murthy
Sheida Nabavi
Sahand Negahban
Kenric Nelson
Daniel Nevo
Rodrigue Ngueyep
Shaoyang Ning
John Nolan
Julie Novak
Isabel Nowinowski
Jp Onnela
Zhengqing Ouyang
Gabriel De Pace
Aaron Palmer
Georgia Papadogeorgou
Chris Parks
Thelge Peiris
Liang Peng
Kendra Plourde
Renato Polimanti
Aleksey Polunchenko
Jean Pouget-Abadie
Vladimir Pozdnyakov
Zhanna Pozdnyakova
Gavino Puggioni
Qi Qi
Jing Qian
Harrison Quick
Eugene Quinn
Vasanthan Raghavan
John Ragland
Sanguthavar Rajasekaran
Pedro Ramos
Nalini Ravishanker
Evan Ray
Byungtae Seo
Nanshi Sha
Zuofeng Shang
Nan Shao
Heather Shappell
Yiyuan She
Daoyuan Shi
Wei Shi
Yulia Sidi
Daniel Smith
Adina Soaita
Seongho Song
Roland Stark
Dale Steele
Matthias Steinruecken
Marilyn Stolar
Jiangwen Sun
Zhaonan Sun
Zhichao Sun
Vasin Suntayodom
Daniel Sussman
Cheng Yong Tang
Jinxin Tao
Mario Tedeschi
Chi-Hse Teng
Zi Han (Sophie) Tew
Suzanne Thornton
Naitee Ting
Elizabeth Upton
Michael Urbanski
Bella Vakulenko-Lagun
Gregory Vaughan
Bryan Vinyard
Rick Vitale
Alexander Volfovsky
Dorothy Wakefield
Phyllis Wan
Wenjie Wang
Tiandong Wang
Yuxiao Wang
Bushi Wang
Fangfang Wang
Haiyang Wang
Zhe Wang
Dong Wang
Biju Wang
Pei Wang
Xia Wang
Haoyu Wang
Joshua Warren
Zheng Wei
Michael Weir
Benedict Wong
Jing Wu
Tingyang Xu
Mengyu Xu
Yanxun Xu
Yishu Xue
Dan Yang
Emmanuel Yashchin
Tairan Ye
Wai-Ki Yip
Zhenkui Zhang
Ting Zhang
Yaohua Zhang
Panpan Zhang
Ying-Ying Zhang
Yuping Zhang
Jiwei Zhang
Chen Zhang
Zhihao Zhang
Yaohua Zhang
Qihao Zhang
Yiming Zhang
Di Zheng
Yan Zhuang
Beth Ziniti
Jingjing Zou
Jian Zou