The 31st New England Statistics Symposium

New England Statistics Symposium

NESS

University of Connecticut

Storrs, Connecticut

April 21 – 22, 2017

NESS 2017 Program

Contents

Welcoming Remarks

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, and 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 was the former Coordinating

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
2. **Vasanthan Raghavan**, Qualcomm Flarion Technologies, New Jersey
3. **Zuofeng Shang**, Binghamton University
4. **Emmanuel Yashchin**, IBM

Oak Hall 235

2. Non-Clinical in Pharmaceutical Industry

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

1. **Don Bennett**, Pfizer
2. **Jerry Lewis**, Biogen
3. **Ray Liu**, Takeda
4. **Chi-Hse Teng**, Novartis

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, Regulatome and Multivariate Analysis

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

1. **Forrest W. Crawford**, Yale
   "Causal Inference for Network Epidemics"
2. **Zhengqing Ouyang**, Jackson Labs
3. **Sijian Wang**, University of Wisconsin Madison
4. **Kuang-Yao Lee**, Yale
   "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. **Pengsheng Ji**, University of Georgia
   "Flexible Spectral Methods for Community Detection"
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**, Floriday 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
   "Pearsons Chi-Squared Statistics: Approximation Theory and Beyond"
2. **Kun Chen**, UConn
   "Robust Dimension Reduction of Correlated Multivariate Data"
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. Survival Analysis

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

1. **Daniel Nevo**, Harvard
2. **Bella Vakulenko-Lagun**, Harvard
3. **Jing Qian**, UMass
4. **Sangwook Kang**

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 Degree Counts in the Preferential Attachment Network"

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**, University of Connecticut
3. **Yasuo Amemiya**, IBM T. J. Watson Research Center

4. **Minge Xie**

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. **Kun Chen**, 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
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.
4. **Nan Shao**, New York Life Insurance
   "Statistical Modeling in the Life Insurance Industry"

Oak Hall 268

4. Biopharmaceutical Session

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

1. **Abidemi Adeniji**, EMD Serono
2. **Bushi Wang**
3. **Joseph c Cappelleri**, Pfizer
   "Meta-Analysis of Safety Data in Clinical Trials"
4. **Qiqi Deng**, Boehringer Ingelheim
5. **Birol Emir**, Pfizer

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
   "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, UMASS
   "Inference from Link-Tracing Network Samples"
4. **Xizhen Cai**, Temple
   "Variable Selection for Dynamic Networks"
5. **Xuan Bi**, Department of Biostatistics, Yale University
   "Genome-Wide Mediation Analysis of Psychiatric and Cognitive Traits in the Philadelphia Neurodevelopmental Cohort"

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 Heart Disease 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

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
   "Assessment of the Adequacy of Asymptotic Theory in Statistical Inference"
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

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

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**, Penn State
   "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
   "Restricted Strong Convexity Implies Weak Sub-Modularity"
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

NESS 2017 Committees

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).

- **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.

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.

- **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.

5. Big Data

- **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 computa-

tional 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.

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.

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.

10. Survival Analysis

- **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.

- **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.

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.

- **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.

12. Feinberg Memorial Session: Bayesian Statistics with Applications

- **DIlli Bhatta**, University of South Carolina Upstate
  "A Bayesian Test of Independence in a Two-Way Contingency Table Under Two-Stage Cluster Sampling with Covariates"

DIlli Bhatta, Balgobin Nandram, Joseph Sedransk

We consider a Bayesian approach for the test of independence to study the association between two categorical variables with covariates using data from a two-stage cluster sampling design. Under this approach, we convert the cluster sample with covariates into an equivalent simple random sample without covariates which provides a surrogate of the original sample. Then, this surrogate sample is used to compute the Bayes factor to make an inference about independence. We apply our methodology to the data from the Trend in International Mathematics and Science Study (2007) for fourth grade U.S. students to assess the association between the mathematics and science scores represented as categorical variables. We show that if there is strong association between two categorical variables, there is no significant difference between the tests with and without the covariates. We also performed a simulation study to further understand the effect of covariates in various situations. We found that in borderline cases (moderate association between the two categorical variables) there are noticeable differences in the test with and without covariates.

Afternoon sessions

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

- **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.

- **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

- **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. Biopharmaceutical Session

- **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.

- **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.

5. Complex Data/Network Modeling

- **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.

- **henry7**, mhlinderm
  "hlin"
  mlkj

  ;L;

6. Spatial Analysis of Public Health Data

- **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.

- **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.

- **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.

8. Statistical Approaches to Data Modeling and 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.

- **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.

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.

- **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

- **Wai-Ki Yip**, Foundation Medicine, Inc.
  "Sr. Biostatistician"
  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.

- **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.

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 un-

biased 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 meta-data. 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 incorporate 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, Elizabeth Schifano, Jun Yan

When large amounts of survival data arrive in streams, conventional estimation methods are difficult to implement due to the requirement of storing all the risk sets up to each accumulation point. In this paper, we apply the cumulative estimating equation(CEE) and cumulatively updating estimating equation(CUEE) to estimate the parameters for the Cox proportional hazards model. Also, to ensure that the proportional hazards assumption holds over time so that all estimators are valid, we propose an online-updating test statistic as well as its variations for the proportional hazards assumption. Their reference distributions are also derived. In simulation studies and real data applications, the test statistic effectively identifies the failure of assumptions and is computationally efficient.

- **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.

NESS 2017 Participants