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**

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**

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**

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**

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**

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**

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**

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**

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**

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**

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**

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**

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**

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**

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**

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 and 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**

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**

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**

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**

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**

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**

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**

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**

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.

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

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.

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.

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.

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

NESS 2017 Participants