The 27th New England Statistics Symposium

Saturday, April 27, 2013
The University of Connecticut
Storrs, Connecticut
Welcoming Remarks

It is my great pleasure to welcome you to the 27th New England Statistics Symposium. On behalf of the Department of Statistics at the University of Connecticut, I would like to wish all the delegates of the 27th New England Statistics Symposium (NESS) an enjoyable visit to our department and the campus. It is hard to believe that 27 years have passed since we have hosted the 1st NESS here at UConn in 1987, with the enthusiastic support from the academic institutions in New England and our colleague, Professor Herman Chernoff, Harvard University. We started our first New England Statistics Symposium with about 50 participants and now the participation has grown to about 200 from academic institutions, industry and government agencies, from all across New England and other states in the US as well. In the past keynote lectures were presented by distinguished speakers within New England, while recently the speakers came from all over U.S. We are thrilled to have this year two distinguished keynote speakers: our own Board of Trustees Distinguished Professor, Richard Bass, University of Connecticut and Arts and Sciences Distinguished Professor, Mike West, Duke University.

This year we offer in conjunction with NESS an extensive educational program, which includes three one-day short-courses on major areas of current interest in statistical science. These short courses are being offered on Friday, April 26, 8:30am 5pm, at the Student Union, University of Connecticut. Professor Mike West, Arts and Science Professor, Department of Statistical Science, Duke University, will offer a course entitled: Bayesian Dynamic Models, Time Series Analysis and Forecasting. Professor Ofer Harel, Department of Statistics, University of Connecticut, and Dr. Gregory Matthews, School of Public Health, University of Massachusetts, will offer a course entitled: An Introduction to the Analysis of Incomplete Data. Professor Jun Yan, Department of Statistics, University of Connecticut, and Dr. Marcos Prates, Department of Statistics, Universidad Federal de Minas Gerais, Brazil, will offer a course on Statistical Analysis of Spatial Data and Visualization with Google Map. I would like to thank the instructors for offering these exciting short courses, which will enrich the knowledge of graduate students and statistical scientists as well.

My sincere thanks go to the leadership of Professors Ming-Hui Chen (Chair), Ofer Harel and Jun Yan for organizing this symposium. I further thank Ms. Megan Petsa, Ms. Tracy Burke and many graduate students who helped at each stage of the planning for this symposium. I am confident that the scientific program for this symposium will be of high quality and lead to the development of exchange of ideas and fruitful interactions between the statistical scientists and graduate students throughout New England and the rest of the country.

I would like to take this opportunity and acknowledge our generous sponsors: UConn Office of Sponsored Programs, Deans Office, College of Liberal Arts and Sciences, UConn Department of Statistics, Connecticut Chapter of the American Statistical Association, Boehringer Ingelheim Pharmaceuticals, Inc., CIGNA, IBM T.J. Watson Research Center, ISBA/EFaB, Pfizer Inc., and
Travelers Insurance Company. I would like to thank the IBM T.J. Watson Research Center, for sponsoring the IBM Student Award for the presentations by graduate students at the NESS. This program initiated in 2005, has attracted this year a record of 20 submissions. I would like to thank the members of the 27th NESS IBM Student Award committee: Professors Ming-Hui Chen (chair), Zhiyi Chi, Ofer Harel, Lynn Kuo and Jun Yan, University of Connecticut and Dr. Fei W. Liu, IBM T.J. Watson Research Center, for their hard work in evaluating the submitted presentations.

I would like to share with you some recent news and accomplishments of our department. Our department was founded in 1963, and is currently one of the largest and major Statistics Departments in New England. Last year on November 1-3, 2013, we have celebrated the 50th Anniversary of the Department. We were fortunate to have in attendance the founder of our department and its first Department Head, Professor Robert H. Riffenburgh, San Diego State University, to present a plenary lecture: The Birth of the UConn Statistics Department: From Clay to Sculpture. During this celebration we have enjoyed several plenary lectures presented by distinguished speakers: Professors Lawrence D. Brown, University of Pennsylvania, Alan E. Gelfand, Duke University and Pranab K. Sen, University of North Carolina. Jointly with the 50th Anniversary Celebration of the Department, we have held three important events. The first event being, the 23rd Pfizer Colloquium, co-sponsored by Pfizer, Inc and ASA. The colloquium lecture entitled: A Pedestrians Lost Horizon in the Wiener Wald of Statistical Science, was presented by Professor Pranab K. Sen, and was followed by a conversation with distinguished statisticians, Professors Malay Ghosh, University of Florida, and Dr. George W. Williams, Vice President, Biostatistics and Epidemiology, AMGEN, in memory of Professor Harry Posten. Both the colloquium lecture and the discussion have been filmed for the Archives of ASA. The second event has been the launching of Makuch Visiting Lecture Series in Biostatistics, sponsored by Professor Robert W. Makuch, Yale School of Public Health, Yale University. The inaugural lecture in this series, entitled: Current Issues in Clinical Trials in Biostatistics, has been delivered by Professor Robert W. Makuch. The third event in conjunction with our 50th Anniversary Celebration has been the presentation of our Departments Distinguished Alumnus Award to Professor E. Jacquelin Dietz, Meredith College. She presented a plenary lecture: JSE, AP Statistics, and Randomization Tests in Stat 101 30 Years of Change in Statistics Education. Our faculty research encompasses all major areas of statistics, including: Applied Probability, Bayesian Analysis, Bioinformatics and Genomics, Biostatistics, Financial statistics, Time Series Analysis and various aspects of Mathematical Statistics and Probability Theory. I am proud to mention that 7 of our current faculty, one emeritus professor, are fellows of the ASA, 7 of our current faculty and one emeritus faculty are fellows of the IMS, and 3 of our current faculty are elected members of the Connecticut Academy of Arts and Sciences or Sciences and Engineering. Many of our faculty received intramural and extramural awards in research excellence. Most of our faculties are editors or editorial board members of major journals. Our graduate and undergraduate programs continue to expand. Currently we have about 90
graduate students and close to 100 undergraduate students majoring in statistics. We have started an MS program with concentration in Biostatistics and are planning to start a Professional MS degree in Biostatistics.

The research initiatives and the quality of output of the department continue to soar. We enjoy research funding from a variety of sources including NSF, NIH and private companies. International and national visibility of the department also continues to grow with our faculty’s participation and visits at conferences and other universities all over the world. We have also developed a strong interdisciplinary research program within UConn. Recently we have received high ranks in NRC as well as US News and World News.

Welcome to UConn and Storrs, the basketball capital of the world!

With best wishes,

Joseph Glaz, Professor and Head
Keynote Speakers

Mike West  
Department of Statistical Science  
Duke University

Mike West is the distinguished Arts and Sciences Professor of Statistical Science in the Department of Statistical Science at Duke University, where he led the development of statistics from 1990-2002. A past president of the International Society for Bayesian Analysis, West has served the international statistics profession broadly. He was involved in the establishment and on founding boards of the National Institute of Statistical Sciences and the Statistical and Applied Mathematical Sciences Institute in the USA, as well as on boards of other national research institutes including the Institute of Statistical Mathematics in Japan, and the Institute for Mathematical Sciences in UK.

West's expertise spans a wide range of areas in Bayesian statistical modelling and computation, focusing on methodology and applications of complex stochastic modelling in high-dimensional problems. His current research emphasizes emerging large-scale computation, multivariate time series and dynamic modelling, sparsity modelling in statistics, and spatio-dynamic modelling. Current areas of emphasis in applications include financial econometrics and societally relevant big data issues in global financial systems, emerging issues of monitoring and fast data analysis in massive dynamic networks in financial and public policy studies, advanced imaging and statistical modelling in systems biology applications, large-scale environmental and atmospheric monitoring, and others.

Since his PhD in 1982, West has published over 160 papers in statistics and interdisciplinary applications in signal processing, finance, econometrics, climatology, public health, genomics, immunology, neurophysiology, systems biology and other areas. He is also co-author of three books on Bayesian time series analysis and dynamic modelling, several edited volumes, and multiple software packages. West has been a statistical consultant for multiple companies, banks, government agencies and academic centers, and was co-founder of a successful biotech company in USA. West's professional kicks come from a variety of sources, but teaching and working with smart, young, emerging statisticians from all over the world is what he prizes most. In addition to working with many undergraduates and Masters students, West has advised over 50 primary PhD students and postdoctoral associates, most of whom are now in faculty, industrial or governmental positions involving advanced statistical research.
Keynote Lecture: Bayesian Dynamic Modelling: Multivariate Time Series, Sparsity and Dynamic Networks

Across many fields of application, the ability to generate an record time series data in increasing dimensions challenges our ability to model, analyse, forecast and effectively utilize resulting data. The challenges and imperatives include a need to increase parsimony in model parametrisations, especially in the large spaces inhabited by time-varying parameter processes of dynamic models. One theme in multivariate dynamic modelling that addresses this challenge is that of dynamic latent thresholding. In recent work, model structures based on this concept have been demonstrably valuable. Applications include studies in finance and econometrics, where relevant latent thresholding structures can result in improved forecasts, decisions and model interpretations, while the ideas are relevant in many other areas of application. This talk will discuss the general idea of latent thresholding as an approach to “dynamic sparsity” modelling, and adaptive, time-varying variable inclusion/selection in time series. This will cover estimation and prediction in several model contexts: dynamic regressions, time-varying vector autoregressions, dynamic latent factor models, and multivariate volatility models. Substantive examples in macroeconomic and financial time series studies will highlight some of the practical interest in the approach. Some of the current interest lies in applications to studies of ”dynamic networks”, where the approach can naturally model time-variation in the structural nature of feed-forward/back interconnections between network nodes. I will highlight these ideas and opportunities in examples in both financial econometrics and experimental neuroscience.
Richard Bass
University of Connecticut

Richard Bass received his Ph.D. from the University of California, Berkeley in 1977 (adviser: P.W. Millar). He was an Assistant Professor, Associate Professor, and Full Professor at the University of Washington (Seattle) (1977-1998). He arrived at the University of Connecticut in 1998 and since 2008 has been a Board of Trustees Distinguished Professor. He has visited Univ. of Illinois, Univ. of Provence, Univ. Paris VI, Univ. of Cambridge, Imperial College, and Univ. of British Columbia (PIMS Distinguished Chair). He is a Fellow of the IMS, a Fellow of the AMS, and has been the editor of the Electronic J. of Probability and of the Transactions of the AMS. He was an invited speaker at the International Congress of Mathematicians in 1994.

Keynote Lecture: A Central Limit Theorem for Symmetric Markov Chains

Consider a symmetric Markov chain on the $d$-dimensional integer lattice. Such a Markov chain is characterized by the electrical conductances between points, and we allow the chain to make long range jumps as well as to move to one of its nearest neighbors. Our main result is that when properly normalized, these Markov chains converge to a Markov process on $d$-dimensional Euclidean space. Depending on the conductances, the limit process could have jumps. Under appropriate conditions, the limit will be a continuous Markov process corresponding to elliptic operators in divergence form. Along the way we will look at transition probability estimates, Harnack inequalities, and harmonic functions. This is joint work with Takashi Kumagai and Toshihiro Uemura.
Schedule

9:00-9:30  Registration & Refreshment

9:30-9:45  Welcome and Opening Remarks, AUST 108
Ming-Hui Chen
Chair of the 27th NESS Planning Committee

Eric Kolaczyk
Department of Mathematics and Statistics
Boston University

Joseph Glaz
Professor and Department Head
University of Connecticut

Jeremy Teitelbaum
Dean of CLAS and Professor of Mathematics
University of Connecticut

Mun Choi
Provost and Executive Vice President for Academic Affairs
University of Connecticut

9:45-10:45  Keynote Presentation, AUST 108

Bayesian Dynamic Modelling: Multivariate Time Series, Sparsity and Dynamic Networks
Mike West
Duke University

Introduction by Nalini Ravishanker, University of Connecticut

10:45-11:00  Break

11:00-12:30  Session 1: Career Opportunities in Statistics
(Panel Discussion)
AUST 202

Session 2: Statistics in Insurance and Actuarial Science (Invited Session)
AUST 102
11:00-12:30  **Session 3:** New Methodologies from Young Researchers (Invited Session)  
AUST 103

**Session 4:** Statistical Inference for Genomics and Stochastics Models (Invited Session)  
AUST 105

**Session 5:** Recent Developments in Spatial Statistics (Invited Session)  
AUST 163

**Session 6:** Statistical Methods for Big Health Science Data (Invited Session)  
AUST 110

**Session 7:** Contributed Session I  
AUST 164

12:30-2:10  **Poster Session,** AUST 313 and 3rd floor hallway

12:45-2:00  **Lunch – Noether Lounge,** AUST 326

2:15-3:15  **Keynote Presentation,** AUST 108

- **A Central Limit Theorem for Symmetric Markov Chains**  
  Richard Bass  
  University of Connecticut

  Introduction by Rick Vitale, University of Connecticut

3:15-3:45  **Break**

3:45-5:15  **Session 8:** Statistical Practice in Business Analytics (Invited Session)  
AUST 102

**Session 9:** New Developments in the Analysis of Incomplete Data (Invited Session)  
AUST 202
3:45-5:15  **Session 10: Probability and Mathematical Statistics**  
(Invited Session)  
AUST 163

**Session 11: Challenges in Modern Modeling Methods**  
(Invited Session)  
AUST 103

**Session 12: Genome-wide Association Studies, Sparse Regression and ODE Inference (Invited Session)**  
AUST 105

**Session 13: Methods for Clinical Trials for Biosimilars (Invited Session)**  
AUST 110

**Session 14: Contributed Session II**  
AUST 164

5:15-6:00  **IBM Student Paper Award Ceremony and Reception**  
Noether Lounge, AUST 326

Dipak K. Dey  
Associate Dean of CLAS, University of Connecticut

J. R. M. Hosking  
IBM T. J. Watson Research Center

Ming-Hui Chen  
Chair, the 27th NESS IBM Student Award Committee  
(Zhiyi, Chi, Ofer Harel, Lynn Kuo, Fei W. Liu and Jun Yan)

6:30-8:30  **Dinner at Chang’s Garden Restaurant**  
1244 Storrs Road, Storrs Commons - Route 195
Detailed Program

Session 1: Career Opportunities in Statistics
Time and Location: 11:00 AM -12:30 PM at AUST 202
Organizer and Moderator: Nai-tee Ting, Boehringer Ingelheim Pharmaceuticals, Inc.

In this session, each panelist will speak for about 5 to 10 min about how to prepare CV, how to handle phone screen, and interview. After that, the session is open for Q and A.

Panelists:
Dominique Haughton, Bentley University
Torey Vizzo, Valesta Clinical Research Solutions
Christopher Parks, Travelers
Jeffrey Young, Cigna

Session 2: Statistics in Insurance and Actuarial Science
Time and Location: 11:00 AM -12:30 PM at AUST 102
Organizers: Emiliano Valdez and Jun Yan, University of Connecticut
Chair: Emiliano Valdez, University of Connecticut

Talks:
   Hua Chen, Temple University
2. Enabling Fact-based Decisions through Statistics
   Javier Abalo, Cigna Healthcare
   Edward Perry, University of Connecticut
4. Predictive Modeling Trends and Implications
   Brian M. Stoll, Towers Watson

Session 3: New Methodologies from Young Researchers
Time and Location: 11:00 AM -12:30 PM at AUST 103
Organizer and Chair: Nitis Mukhopadhyay, University of Connecticut

Talks:
1. A Marginalization Approach to High Dimensional Covariance Matrix Estimation
   Samprit Banerjee, Cornell University
2. Inference for the Broken-Stick Model: a Computationally Faster Approach
   Ritabrata Das, University of Michigan

3. Multiple Crossing Sequential Fixed-Size Confidence Region Methodologies for Normal Mean Vector
   Sankha Muthu Poruthotage, University of Connecticut

Session 4: Statistical Inference for Genomics and Stochastic Models
Time and Location: 11:00 AM -12:30 PM at AUST 105
Organizer and Chair: Luis Carvalho, Boston University

Talks:

1. Maximum Likelihood Estimation for Diffusion Models with Multiple Scales and Small Noise
   Konstantinos Spiliopoulos, Boston University

2. On Estimation and Prediction Problems for Stationary Time Series Models with Memory
   Mamikon Ginovyan, Boston University

3. Graph-regularized Centroid Estimation on a Hierarchical Bayesian Model for Genome-wide Association Studies
   Luis Carvalho, Boston University

4. Robust Sensitivity Analysis for Stochastic Systems
   Henry Lam, Boston University

Session 5: Recent Developments in Spatial Statistics
Time and Location: 11:00 AM -12:30 PM at AUST 163
Organizer: Luke Bornn, Harvard University
Chair: Zhiyi Chi, University of Connecticut

Talks:

1. Nonstationary Modeling through Dimension Expansion
   Luke Bornn, Harvard University

2. Spatial Models for Traffic Related Air Pollution
   Theodore R. Holford, Yale University

3. Bayesian Spline Models for the Analysis of Spatio-Temporal Count Data
   Cici Bauer, Brown University
Session 6: Statistical Methods for Big Health Science Data
Time and Location: 11:00 AM -12:30 PM at AUST 110
Organizer: Xihong Lin, Harvard University
Chair: Elizabeth Schifano, University of Connecticut

Talks:
1. Attribute-dependent Partial Correlation Estimation for Gaussian Graphical Models
   Jian Guo, Harvard University
2. Using Medicare Claims to Study Patient-sharing Networks of Physicians
   JP Onnela, Harvard University
3. Cross-Study Validation of Prediction Methods: a Bayesian Nonparametric Analysis
   Lorenzo Trippa, Harvard University

Session 7: Contributed Session I
Time and Location: 11:00 AM -12:30 PM at AUST 164
Chair: Sangwook Kang, University of Connecticut

Talks:
1. Investigating the Effect of Environmental Factors on Phytoplankton Count
   Kalyan Das, Calcutta University
2. Homicide Rate Analysis with Dynamic Models for Near-Gaussian Random Fields Using INLA
   Marcos O. Prates, Universidade Federal de Minas Gerais
3. Network-guided Sparse Regression Modeling for Detection of Gene by Gene Interactions
   Chen Lu, Boston University
4. On Quinns Mathematics Is Not Science: Implications for Our Statistical Methodologies and Mathematical Biology
   G. Arthur Mihram, Princeton, NJ 09542-1188
5. An Adaptive Procedure for Multiple Window Scan Statistics
   Tung-Lung Wu, University of Connecticut
Session 8: Statistical Practice in Business Analytics

Time and Location: 3:45 PM - 5:15 PM at AUST 102

Organizer: Fei W. Liu, IBM Thomas J. Watson Research Center
Chair: J. R. M. Hosking, IBM Thomas J. Watson Research Center

Talks:

1. A Hierarchical State-space Model for Short-term Forecasting of Residential Electricity Demand
   J. R. M. Hosking, IBM T. J. Watson Research Center

2. NOTAM²: Nonparametric Bayes Multi-Task Multi-View Learning
   Hongxia Yang, IBM T.J. Watson Research Center

3. Optimal Selection of Building Components Using Sequential Design via Statistical Surrogate Models
   Rui Zhang, IBM T.J. Watson Research Center

4. Statistical Modeling of Ambient Air Quality for IBM Smarter Planet Initiative
   Youngdeok Hwang, IBM T. J. Watson Research Center

Session 9: New Developments in the Analysis of Incomplete Data

Time and Location: 3:45 PM - 5:15 PM at AUST 202

Organizer and Chair: Ofer Harel, University of Connecticut

Talks:

1. Strategy for Intention to Treat Analysis in Randomised Trials with Missing Outcome Data
   Nicholas Horton, Smith College

2. Non-ignorable Missing and Misreported Count Data: Parametric Models and Inference
   Forrest W. Crawford, Yale University

3. Joint Modeling of Incomplete Data with Diverse Variable Types Using Latent-Variable Models
   Thomas R. Belin, University of California, Los Angeles
Session 10: Probability and Mathematical Statistics  
**Time and Location:** 3:45 PM - 5:15 PM at AUST 163  
**Organizer and Chair:** Vladimir Pozdnyakov, University of Connecticut

Talks:

1. Statistical Inference and Optimalities in Estimation of Gaussian Graphical Model  
   **Zhao Ren, Yale University**

2. Density Estimation for Binary Sequences  
   **Karthik Bharath, Ohio State University**

3. Stochastic Fixed-point Equations Arising in the Analysis of Algorithms  
   **Mariana Olvera-Cravioto, Columbia University**

Session 11: Challenges in Modern Modeling Methods  
**Time and Location:** 3:45 PM - 5:15 PM at AUST 103  
**Organizer and Chair:** Krista Gile, University of Massachusetts Amherst

Talks:

1. On the Uncertainty Evaluation of the Best Linear Unbiased Predictor  
   **Rongheng Lin, University of Massachusetts Amherst**

2. Determination of Proportionality in Two-part Models and Analysis of Multi-Ethnic Study of Atherosclerosis (MESA)  
   **Anna Liu, University of Massachusetts Amherst**

3. Assessment of the Number of Components in Gaussian Mixture Models in the Presence of Multiple Local Maximizers  
   **Daeyoung Kim, University of Massachusetts Amherst**

4. Adjustment of Dependent Truncation with Inverse Probability of Weighting  
   **Jing Qian, University of Massachusetts, Amherst**

5. Bayesian Inference with Identifiable Subjects in the Prior  
   **Edward J. Stanek III, University of Massachusetts Amherst**

Session 12: Genome-wide Association Studies, Sparse Regression and ODE Inference  
**Time and Location:** 3:45 PM - 5:15 PM at AUST 105  
**Organizer:** Hongyu Zhao, Yale University  
**Chair:** Lynn Kuo, University of Connecticut

Talks:

1. Parameter Estimation in ODE by Iterative Integration  
   **Xin Qi, Georgia State University**
2. Truncated and Weighted Z-transformation Methods for Combining P-values
   **Zheyang Wu**, *Worcester Polytechnic Institute*

3. Sparse Regression by Projection
   **Ruiyan Luo**, *Georgia State University*

4. Leveraging Local IBD Increases the Power of Case/Control GWAS with Related Individuals
   **Joshua Neil Sampson**, *National Cancer Institute*

**Session 13: Methods for Clinical Trials for Biosimilars**
**Time and Location**: 3:45 PM - 5:15 PM at AUST 110
**Organizer and Chair**: Gheorghe Doros, Boston University

**Talks:**

1. Covariate Effect on Constancy Assumption in Non-inferiority Clinical Trials
   **Siyan Xu**, *Boston University*

2. How to Define a Margin When the 95-95 Rule Does Not Work and How to Analyze (Sometimes You Have to Cram a Square Peg into a Round Hole)
   **Kerry Barker**, *Pfizer Inc.*

3. Totality Test and Adaptive Design for Bio-similarity
   **Mark Chang**, *AMAG Pharmaceutical and Boston University*

4. Blinded Sample Size Re-Estimation Using an Adaptive Margin in Non-Inferiority Clinical Trials with a Binary Endpoint
   **Joe Massaro**, *Boston University*

**Session 14: Contributed Session II**
**Time and Location**: 3:45 PM - 5:15 PM at AUST 164
**Chair**: Xiaojing Wang, University of Connecticut

**Talks:**

1. Intrusion as (Anti)social Communication: Characterization and Detection
   **Natallia Katenka**, *University of Rhode Island*

2. Genetic Macrostructure, Replicability: What Can We Learn from the Di-allel?
   **Alan Lenarcic**, *Harvard University, Jackson Labs, University of North Carolina Chapel Hill*

3. GEE-Based Intervals for the Difference Between Two Treatment Means in Clinical Trials
   **Krishna K. Saha**, *Central CT State University*
4. A Versatile Model to Analyze Life Time Data

Jie Huang, University of Maine

Session 15: Poster Session

Time and Location: 12:30 PM - 2:10 PM at AUST 313 and 3rd floor hallway

Posters:

1. Enhanced Archaeological Predictive Modelling in Space Archaeology: a Statistical Learning Approach on Remotely Sensed Imageries for Identifying Archaeological Sites
   Li Chen, Johns Hopkins University

2. Nonparametric Bayesian Inference for the Number of Components in a Finite Mixture
   Jeffrey W Miller, Brown University

3. A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data
   Nicholas A. James, Cornell University

4. Impact of Prior Distribution Uncertainty in Multiple Imputation Inference
   Valerie L. Pare, University of Connecticut

5. Dynamic Compositional Modeling of Pedestrian Crash Counts on Urban Roads in Connecticut
   Volodymyr Serhiyenko, University of Connecticut

6. Fitting Large-scale GLMs with Implicit Updates
   Panos Toulis, Harvard University

7. Utilizing Protein Structure to Identify Non-Random Somatic Mutations
   Gregory Ryslik, Yale University

8. Handling Data with Three Types of Missing Values
   Jennifer Boyko, University of Connecticut

9. Integrative Analysis of Prognosis Data on Multiple Cancer Subtypes Using Compound Group Bridge
   Jin Liu, Yale University

    Dahlia Nadkarni, Brown University

11. Comparing Regression Coefficients Using Incomplete Data
    Chantal Larose, University of Connecticut

12. Integrative Analysis of Cancer Prognosis Data With Contrast Group Bridge Penalization
    Xingjie Shi, Shanghai University of Finance and Economics, Yale University
13. Estimating the Average Treatment Effects of Nutritional Label Use Using Subclassification with Regression Adjustment
   **Michael J. Lopez, Brown University**

14. A Bayesian Test of Independence in a Two-way Contingency Tables Under Cluster Sampling
   **Dilli Bhatta, Worcester Polytechnic Institute**

15. Population Based Convergence Criterion for Self-Organizing Maps
   **Benjamin Ott, University of Rhode Island**

   **Hossein Azari Soufiani, Harvard University**

17. Bayesian Predictive Model Assessment and Selection with Bregman Divergence
   **Gyuhyeong Goh, University of Connecticut**

18. A Structured Covariance Model for Quantifying Transcriptional Control of Protein Levels
   **Alexander Franks, Harvard University**

19. Approximating the Convolution of Lognormal Random Variables within a MCMC Sampler
   **Guillaume Basse, Harvard University**

20. Bayesian Degree-corrected Stochastic Block Models for Community Detection
   **Lijun Peng, Boston University**

21. Incorporating Spatial Dependence in Regional Frequency Analysis
   **Zhuo Wang, University of Connecticut**

22. On the Characterization of a Class OfFisher-Consistent Loss Functions and Its Application to Boosting for Hierarchical Outcomes
   **Matey Neykov, Harvard University**

23. An EM Algorithm for the Matrix Normal Distribution
   **Hunter Glanz, Boston University**

24. Estimating Network Degree Distributions Under Sampling: an Inverse Problem, with Applications to Monitoring Social Media Networks
   **Yaonan Zhang, Boston University**

   **Harold Bae, Boston University**

26. F-tests with Incomplete Data Under the Multiple Regression Setup
   **Ashok Chaurasia, University of Connecticut**
27. Cause & Effect: Standardizing Knowledge through Statistics
   Mike Sagherian, University of Connecticut

28. Local Algorithm to Estimate Graphons by Stochastic Blockmodels Approximation
   Thiago Costa, Harvard University

   Jingshu Liu, Boston University

30. Inference Based on Estimating Equations for Logarithmic Autoregressive
    Conditional Duration (ACD) Models
   James Anderson, University of Connecticut

31. Development of Healthy Eating Preference Index: Evidence of Content
    and Predictive Validity
   Mastaneh Sharafi, University of Connecticut

32. Defect Growth Prediction in Rolling Bearings Based on Approximate Entropy
   Peng Wang, University of Connecticut

33. A Noise-assisted Data Processing Method in Weak Signal Detection: Stochas-
    tic Resonance and Its Adaptive Scheme
   Rui Zhao, University of Connecticut

34. Spatially Weighting Genes in Variable Selection for Genome-Wide Asso-
    ciation Studies
   Ian Johnston, Boston University

35. The Network You Keep: Graphlet-Based Discrimination of Persons of
    Interest
   Saber Shokat Fadaee, Northeastern University

36. Behavioral Versus Biological Measures on Sexual Risk: a Systematic Re-
    view and a Comparison of Intervention Effect Sizes
   Samantha Russo, University of Connecticut

37. Conditional Inference for Edge Covariate Effects in Sparse Networks
   Daniel Klein, Brown University
Abstracts of Invited Papers

A Marginalization Approach to High Dimensional Covariance Matrix Estimation

Samprit Banerjee, Cornell University
Stefano Monni, German Cancer Research Center

Estimation of the covariance matrix, especially in higher dimensions is a challenging statistical problem which is of great interest in many applications especially in genetics and genomics. The existing methods of estimation (both Bayesian and non-Bayesian) pose a number of challenges even when the dimension of the covariance matrix is moderately large and in such situations the usual solution is to assume a structure for the true covariance matrix. However, in reality such constraints might not be meaningful. In this paper estimation of the unconstrained covariance matrix is considered. An orthogonally invariant estimator is proposed by considering the marginal distribution of eigenvalues. The marginal distribution is analytically intractable, hence an approximation for the likelihood function is considered, maximum likelihood estimators provided, and its properties investigated. Finally, a non-informative prior is proposed and performance of the corresponding estimator is investigated via simulations.

Inference for the Broken-Stick Model: a Computationally Faster Approach

Ritabrata Das, University of Michigan
Moulinath Banerjee, University of Michigan
Bin Nan, University of Michigan

The existence of one or more change-points in linear regression problems has significant applications in climate data, economic time series and for modeling biological processes, where the change-points mostly pertain to the onset of biologically important phenomena. Estimation of change-point(s) in a broken-stick model using the exact likelihood has been discussed in some depth in the literature but most of the methods are computationally quite expensive: the non-differentiability at the kink(s) necessitates an exhaustive search across tuples of order statistics. In this article, we present a smoothing based approach to address this difficulty. We smooth the broken-stick in a shrinking neighborhood of the kinks by quadratic functions and use this as our working model, which allows the use of Newton-Raphson type methods for the working likelihood function. Asymptotic properties of our estimates are presented. We find that our estimates converge at $\sqrt{n}$-rate and are as asymptotically efficient as the exact MLE. Simulations clearly vindicate the computational economy of our approach.
A Hierarchical State-space Model for Short-term Forecasting of Residential Electricity Demand

J. R. M. Hosking, IBM T. J. Watson Research Center
Ramesh Natarajan, IBM T. J. Watson Research Center
Soumyadip Ghosh, IBM T. J. Watson Research Center
Shivaram Subramanian, IBM T. J. Watson Research Center
Xiaoxuan Zhang, IBM T. J. Watson Research Center

Electricity transmission and distribution systems are increasingly moving towards the use of a "Smart Grid", a densely instrumented network for electricity delivery that facilitates remote control, automation, and feedback to consumers. This evolution introduces new challenges in modeling and forecasting electricity usage data. We describe a model of the daily load curve for residential electricity consumption that includes the effects of dynamic price incentives on the demand response, which is a topic of considerable interest in the Smart Grid. The model is based on a time series and regression framework in which the observed values of the daily load curve are represented by a set of periodic smoothing-spline basis functions, with the coefficients of the basis functions evolving according to a linear Gaussian state-space model that incorporates mean shifts, day-of-week and holiday adjustments, and temperature effects, in addition to the dynamic price incentive effects. This modeling and forecasting methodology provides the ability to model intraday load substitution effects that are induced by the specified dynamic pricing schedules, the ability to use fine-grained (5 to 15 minute interval) observational data without greatly increasing the computational cost of the estimation and forecasting procedures, and fast updating of model forecasts as new usage and weather data arrive.

Parameter Estimation in ODE by Iterative Integration

Xin Qi, Georgia State University

A new method of parameter estimation in ODE based on Picards method is proposed. An initial estimate of the solution curve is found by usual non-parametric methods. Then the integration of a functional of the estimate is calculated as an updated estimate. The estimate of parameters is based on the updated estimate of the solution and the nonlinear least square methods.

Truncated and Weighted Z-transformation Methods for Combining P-values

Zheyang Wu, Worcester Polytechnic Institute
Tiejun Tong, Hong Kong Baptist University

In genetics and epigenetics it is critical to develop powerful statistical methods to identify associations between a response and sets of genetic markers or
genes. From biological perspective, due to potential biological joint effects, such as interactions, simultaneously studying a group of markers or genes together has potential to strengthen the association by incorporating such joint effects, and thus improve the power of signal detection. From statistical perspective, due to the threshold effect on the signal strength, when comparing with the identification of individual signals, the identification of some signals existing in a group of candidates has weaker requirements for signal strength and density. To address such problem, hypothesis test statistics are developed based on combining individual p-values. Proper procedures of truncating, weighting, and transforming the p-values can potentially help signals standing out, and thus improve statistical power. In this research, we give a unified study of p-value combination methods by defining a general function of Z-transformation that covers various truncation and weighting procedures. Theoretical results and simulation show that the quadratic function of truncated Z-transformation provide a higher statistical power than other methods in most setups of signal strength and density. The methods are applied and compared in real data analysis of genome-wide association studies.

**Sparse Regression by Projection**

Ruiyan Luo, Georgia State University  
Xin Qi, Georgia State University  
Raymond J Carrol, Texas A&M University  
Hongyu Zhao, Yale University

A new framework, regression by projection and its sparse version, is proposed. Based on the relationship between this framework and discriminant analysis for classification, we also propose a new sparse discriminant analysis method. The novel feature of this method is that in addition to sparsity, zero within-class and between-class correlations among the components can be achieved simultaneously, which benefits prediction. We conduct simulations and empirical data analysis to compare our method with well-established methods such as the LASSO, the elastic net and others. Our results suggest that our new framework has better performance for both regression and discriminant analyses, as well as for variable selection. This is especially true for high-dimensional data and data with a relatively large number of classes.

**Leveraging Local IBD Increases the Power of Case/Control GWAS with Related Individuals**

Joshua Neil Sampson, National Cancer Institute  
Bill Wheeler, Information Management Services  
Peng Li, National Cancer Institute  
Jianxin Shi, National Cancer Institute

Large case/control genome-wide association studies (GWAS) often include
groups of related individuals with known relationships. Our goal is to introduce a new test statistic for GWAS that can (1) accommodate individuals of arbitrary, but known, relationships and (2) offer significantly improved power when there is a large number of family-based controls. This chromosome-based Quasi Likelihood Statistic (cQLS) focuses on founder chromosomes and their alleles, as opposed to subjects and their genotypes, and can therefore leverage information about local Identity By Descent (IBD) between relatives to improve study power. We demonstrate this improvement by comparing cQLS to traditional test statistics using simulated datasets. Moreover, our simulations reevaluate the benefits of genotyping an affected sibling of a study case, concluding that the gain in power exceeds that from genotyping a randomly ascertained case offering, on average, twice as many unique chromosomes. Simulations also demonstrate the diminished benefit of genotyping cases with a family history as GWAS search for smaller effect sizes.

Strategy for Intention to Treat Analysis in Randomised Trials with Missing Outcome Data

Nicholas Horton, Smith College
James Carpenter, London School of Hygiene and Tropical Medicine
Ian White, MRC/Cambridge

Loss to follow-up is often hard to avoid in randomised trials, but is not always appropriately accounted for in analyses. This talk will suggest a framework for intention to treat analysis that depends on making plausible assumptions about the missing data and including all participants in sensitivity analyses. The strategy includes the following principles: 1) Attempt to follow up all randomised participants, even if they withdraw from allocated treatment; 2) Perform a main analysis of all observed data that are valid under a plausible assumption about the missing data; 3) Perform sensitivity analyses to explore the effect of departures from the assumption made in the main analysis; 4) Account for all randomised participants, at least in the sensitivity analyses.

NOTAM²: Nonparametric Bayes Multi-Task Multi-View Learning

Hongxia Yang, IBM T.J. Watson Research Center
Jingrui He, Stevens Institute of Technology

Heterogeneous learning refers to addressing problems with multiple types of heterogeneity, e.g., task heterogeneity, view heterogeneity, etc. It finds abundant applications in cross-lingual document classification, cross-domain sentiment analysis, web image classification, etc. Traditional approaches handle different types of heterogeneity separately via multi-task learning, multi-view learning, etc. More recently, researchers start to jointly model different types of heterogeneity in order to improve the learning performance with limited training data. In this paper, we advance state-of-the-art in heterogeneous learning by
jointly modeling task and view relatedness via nonparametric Bayes method. To be specific, we model task relatedness using normal penalty with sparse covariances to couple multiple tasks and view relatedness using matrix Dirichlet process. We also propose NOTAM^2 algorithm, which is based on an efficient Gibbs algorithm. Experimental results demonstrate the effectiveness of NOTAM^2.

Non-ignorable Missing and Misreported Count Data: Parametric Models and Inference

Forrest W. Crawford, Yale University
Robert E. Weiss, University of California, Los Angeles
Marc A. Suchard, University of California, Los Angeles

Social scientists have developed a nuanced understanding of the ways in which respondents sometimes modify or omit personal information in public health surveys. For example, researchers have characterized patterns of non-response, rounding, estimating, and lying in answers to questions about taboo subjects like sexuality or drug use. In many cases, subjective or non-quantitative ideas about misreporting and non-response can be used to specify models and prior distributions for learning about missing data mechanisms in the analysis of survey results. In this talk, I explore new ways of analyzing apparently heaped or rounded count data from a public health study of the sexual behavior of high-risk HIV positive youth using a Bayesian hierarchical model and a novel class of reporting distributions. This allows joint inference of population- and individual-level parameters and gives valuable insight into the misreporting mechanism.

Optimal Selection of Building Components Using Sequential Design via Statistical Surrogate Models

Rui Zhang, IBM T.J. Watson Research Center
Fei Liu, IBM T.J. Watson Research Center
Angela Schoergendorfer, IBM T.J. Watson Research Center
Youngdeok Hwang, IBM T.J. Watson Research Center
Young M. Lee, IBM T.J. Watson Research Center
Jane L. Snowdon, IBM T.J. Watson Research Center

Choosing the optimal combination of building components that minimize investment and operational costs is a topic of great importance in the building simulation community. Optimization using simulation tools, i.e., EnergyPlus, becomes computationally expensive for traditional search approaches. An additional challenge is the complexity of the input parameter space, which is usually very large and contains both continuous and discrete variables. In this paper, we present a novel approach to address both of these problems. The key idea of the proposed approach is to first build a statistical surrogate model for the energy simulation model and to then update the surrogate model based on the concept
of sequential design of experiments. We demonstrate the proposed approach using a case study of a live retrofit project for Building 661 at the Navy Yard of Philadelphia, USA. Results show that the statistical surrogate model allows for fast evaluation of the building’s energy consumption, and the sequential design reduces the computational cost by requiring a smaller number of runs of the energy simulation model.

**Nonstationary Modeling through Dimension Expansion**

Luke Bornn, *Harvard University*

In this talk, we propose a novel approach to modeling nonstationary spatial fields. The proposed method works by expanding the geographic plane over which these processes evolve into higher dimensional spaces, transforming and clarifying complex patterns in the physical plane. By combining aspects of multi-dimensional scaling, group lasso, and latent variable models, a dimensionally sparse projection is found in which the originally nonstationary field exhibits stationarity. Following a comparison with existing methods in a simulated environment, dimension expansion is studied on a classic test-bed data set historically used to study nonstationary models. Following this, we explore the use of dimension expansion in modeling air pollution in the United Kingdom, a process known to be strongly influenced by rural/urban effects, amongst others, which gives rise to a nonstationary field.

**Spatial Models for Traffic Related Air Pollution**

Theodore R. Holford, *Yale University*

Geographic Information Systems (GIS) provide a valuable tool for studying associations between environmental exposures and health. A commonly used approach is to construct regressor variables from landscape features near a residence, which are then included in traditional regression models. This method often gives equal weight to all features within a buffer and the size of the buffer is often arbitrary. Alternatively, one can use a specified model that has been derived, such as the CALINE model develop in California during the 1970s. The limitation of this approach is that may not work well in other regions and the vehicle fleet has changed considerably in the intervening decades. An alternative approach is presented in which highways are considered to be a locus of points on a line and cumulative exposure results from integration over the line. This dispersion function related to an outcome is directly estimated as a function of distance and bearing of road segments in relation to a point. In addition, it enables one to include the impact of meteorology as well as other landscape features the may modify the cumulative exposure. Results from this approach are readily interpretable in terms of the distribution of chemical agents generated by a pollution source. A model for the effect of traffic on childhood asthma in Connecticut is used to illustrate the method.
Bayesian Spline Models for the Analysis of Spatio-Temporal Count Data

Cici Bauer, Brown University

In this talk I will focus on the development of spatial-temporal models for infectious disease count data. This topic was motivated by surveillance data for hand-foot-mouth disease (HFMD) collected in China between 2009 and 2010. The overall strategy we take is to decompose the log relative risk into three components: a large-scale temporal trend, a large-scale spatial trend and a spatial-temporal interaction. The interaction is modeled using a bivariate tensor product of cubic B-splines. We fit the model in a Bayesian framework and the structure of the interaction between space and time is imposed through a prior on the coefficients of the basis functions which are constructed as a tensor product of cubic B-splines. This model is amenable to prediction through the use of Gaussian Markov Random Field (GMRF) priors.

Statistical Inference and Optimalities in Estimation of Gaussian Graphical Model

Zhao Ren, Yale University
Tingni Sun, University of Pennsylvania
Cun-Hui Zhang, Rutgers University
Harrison H. Zhou, Yale University

Gaussian graphical model has a wide range of applications. The study of Gaussian graphical model had attracted a lot of attention recently. This paper considers a fundamental question: when is it possible to obtain statistical inference for estimation of Gaussian Graphical Model? A novel regression approach is proposed to obtain asymptotically efficient estimation of each entry when the precision matrix is sufficiently sparse. When the precision matrix is not sufficiently sparse, i.e., the sparseness condition fails, a lower bound is established to show that it is no longer possible to achieve the parametric rate estimation of each entry through a novel construction of a subset of sparse precision matrices and Le Cam’s Lemma.

The asymptotic normality result is applied to do adaptive support recovery, to obtain adaptive rate-optimal estimation of the precision matrix under various matrix $l_q$ norms, and to do inference and estimation for latent variable graphical models, without the irrepresentable condition and the $l_1$ constraint of the precision matrix which are commonly required in literature. Numerical results confirm our theoretical findings.
Multiple Crossing Sequential Fixed-Size Confidence Region Methodologies for Normal Mean Vector

Nitis Mukhopadhyay, *University of Connecticut*
Sankha Muthu Poruthotage, *University of Connecticut*

The purely sequential procedure proposed by Mukhopadhyay and Al-Mousawi (1986) can be used to construct a confidence region for the mean vector of $N_p(\mu, \sigma^2 H)$ with known $H_{p \times p}$ but unknown $\sigma^2$. The main advantage of this procedure is that the size of the confidence region ($= d$) and the confidence coefficient ($= 1 - \alpha$) could be fixed in advance. Even though this procedure has efficiency, and asymptotic consistency properties it does not have the consistency property. In this article, a purely sequential sampling procedure is proposed which allows for continuous sampling even after the sample size exceeds the corresponding boundary ($= a \frac{S^2}{d^2}$) multiple times. Hence the stopping rules of this kind are called multiple crossing stopping rules. Some analytical properties such as efficiency, and asymptotic consistency of multiple crossing stopping rules are discussed. Its ability to achieve the required coverage probability without significant over sampling is illustrated via a large scale simulation exercise, across a wide range of values of $d$ and $\alpha$. A truncation technique as well as a fine-tuned adjustment to the multiple crossing stopping rule is proposed which are intended to increase the usability of the proposed procedure. Then we demonstrate how this methodology can be effectively applied for constructing fixed-sized confidence regions for treatment effects of ANOVA models. Finally some real data illustrations are provided in order to demonstrate the actual implementation and possible applications of the proposed procedure.

Determination of Proportionality in Two-part Models and Analysis of Multi-Ethnic Study of Atherosclerosis (MESA)

Anna Liu, *University of Massachusetts Amherst*
R. Kronmal, *University of Washington*
X. Zhou, *University of Washington and Veterans Affairs Puget Sound Health Care System*
S. Ma, *Yale University*

In MESA (Multi-Ethnic Study of Atherosclerosis), it is of interest to model the development and progression of CAC (coronary artery calcium). With about half of the CAC scores equal to zero and the rest continuously distributed, semiparametric two-part models are needed. Our main interest lies in determining the (partial) proportionality between the two covariate effects in two-part models. Such an investigation can provide important information on the mechanisms underlying CAC development. We propose a novel approach, which consists of penalized maximum likelihood estimation and a step-wise hypothesis testing procedure to determine proportionality. Simulation shows satisfactory performance of the proposed approach. Analysis of MESA suggests that proportion-
ality holds for all covariates except LDL and HDL.

**Maximum Likelihood Estimation for Diffusion Models with Multiple Scales and Small Noise**

Konstantinos Spiliopoulos, *Boston University*

We study the problem of parameter estimation for stochastic differential equations with small noise and fast oscillating parameters. Depending on how fast the intensity of the noise goes to zero relative to the parameter that separates the scales, we consider three different regimes. For each regime, we construct the maximum likelihood estimator and we study its consistency and asymptotic normality properties. An important application that also motivated this research is parameter estimation for the first order Langevin equation with a two scale potential in rough energy landscapes. A related simulation study demonstrating the theoretical results will be provided. This is joint work with Alexandra Chronopoulou.

**On Estimation and Prediction Problems for Stationary Time Series Models with Memory**

Mamikon Ginovyan, *Boston University*

During the last decades long-memory and anti-persistent time series models have become an important part of theoretical and applied time series analysis. These models are characterized by slowly decaying correlation functions at infinity, or equivalently, by spectral densities possessing singularities (poles or zeros) at the origin. These features change in an essential way the statistical estimation and prediction procedures, and as a consequence, many of the methods and results used for analyzing short-memory time series models are no longer appropriate for these models. In this talk we will discuss linear prediction and nonparametric estimation problems for stationary random processes that may display long, intermediate, or short memory and seasonality behavior. Let \( \{X(t), t \in T\} \) be a continuous- or discrete-time second-order stationary process possessing a spectral density \( f(\lambda), \lambda \in \Lambda \). The time domain \( T \) is the real line \( \mathbb{R} \) in the continuous-time case, and the set of integers \( \mathbb{Z} \) in the discrete-time case. The frequency domain \( \Lambda \) is \( \mathbb{R} \) in the continuous-time case, and is \( (-\pi, \pi] \) in the discrete-time case. The Linear Prediction Problem: Having observed part of the past, one wishes to predict the future, and to describe the rate of decrease of the relative prediction error to zero as the length of observed past increases. This rate is likely to depend on the dependence structure of the model and the smoothness properties of its spectral density \( f(\lambda) \). The Nonparametric Estimation Problem: Assuming that the spectral density \( f \) of the model is unknown, and belongs to a given class \( \Theta \), construct asymptotically efficient estimators for the value \( \Phi(f) \) of a given functional \( \Phi(\cdot) \) at an unknown point \( f \in \Theta \). The solution depends on the type of functional \( \Phi(f) \) - whether it is linear or non-linear,
on the memory structure of the model $X(t)$, and the smoothness structure of the "parametric" set $\Theta$.

**Graph-regularized Centroid Estimation on a Hierarchical Bayesian Model for Genome-wide Association Studies**

Luis Carvalho, *Boston University*

Genome-wide association studies (GWAS) attempt to determine which genomic markers (SNPs) are predictors of genetic traits, most commonly human diseases. In practice, despite the extreme imbalance of having millions of markers recorded for only a few thousand individuals, it is of great interest to glean as much information as possible from this type of data. To this end, we propose a novel Bayesian statistical model that exploits a hierarchical structure between markers and genes to leverage information between levels and alleviate the "large p small n" regimen while still attaining a reasonably complex and realistic model. We further describe a collapsed Gibbs sampler that takes advantage of particular features of the resulting graphical model to obtain an efficient sampling procedure. We conduct inference on which SNPs and genes are associated with the studied trait using graph-regularized centroid estimation. Finally, we illustrate the proposed model and estimation procedure on simulated data and offer initial results on real-world data. If time permits, we also discuss a latent genotype procedure that aims to correct genotypical correlations. This is joint work with Ian Johnston.

**Assessment of the Number of Components in Gaussian Mixture Models in the Presence of Multiple Local Maximizers**

Daeyoung Kim, *University of Massachusetts Amherst*
Byungtae Seo, *Sungkyunkwan University*

Finite mixtures of Gaussian distributions are attractive in identifying the underlying group structure in the data. However, the likelihood inference for the Gaussian mixtures is theoretically and practically challenging because it is unbounded and often has multiple local maximizers. In particular, the presence of multiple local maximizer including spurious local maximizers affects the performances of the model selection criteria used to choose the number of components. We propose a new type of likelihood-based method, a k-deleted likelihood method, designed to avoid spurious local maximizers and choose a statistically desirable local maximizer in the presence of multiple local maximizers for a Gaussian mixture model. We examine, by a real example and simulation studies, the performance of the proposed method in the likelihood-based model selection criteria commonly used to assess the number of components in Gaussian mixture models.
Robust Sensitivity Analysis for Stochastic Systems

Henry Lam, Boston University

We propose a parameter-free framework for performing sensitivity analysis for large-scale stochastic systems. The analysis uses the Kullback-Leibler divergence as a measure of model discrepancy, and aims to obtain a well-defined notion of derivative estimators. The properties of these estimators and the methodology involved in the derivation will be discussed.

Systemic Risk Measures in the Insurance Industry: a Factor Copula Approach

Hua Chen, Temple University
J. David Cummins, Temple University
Krupa S. Viswanathan, Temple University
Mary A. Weiss, Temple University

This paper uses factor copula models to analyze systemic risk in the U.S. insurance industry based on daily stock return data. The sample consists of forty publicly traded life-health and property-casualty insurers over the sample period 2002-2011. We analyze insurers using four measures of systemic risk that have been proposed in the recent literature, i.e., CoVar, modified CoVar, marginal expected shortfall (MES), and SRISK. The results of the copula models serve as essential inputs into the systemic risk measures. We find that both life-health and property-casualty insurers are susceptible to systemic risk. However, when financial markets are in distress, life-health insurers respond more significantly to the market factor and thus are more risky than property-casualty insurers.

Enabling Fact-based Decisions through Statistics

Javier Abalo, Cigna Healthcare

At Cigna Analytics we focus our efforts in understanding the customer needs and helping our business partners clarify priorities, align resources, and design efficient and effective strategic plans to provide excellent customer value that drive growth and strength. In a simple statement, we enable fact-based decision making by asking questions, addressing problems, and identifying opportunities for the business. How we do that has a very simple answer: mostly, and primarily, through Research and Statistics. We use Statistics to cover our five analytic service lines: Trend Analytics (What happened?), Root Cause Analytics (Why did this happen?), Predictive Analytics (What will happen next?), Opportunity Analytics (Where are the opportunities), and Innovation Analytics (How can we get better?). We provide Specific, Measurable, Actionable, Responsive and Time-bound (SMART) Innovation through Statistics. And we Design, Test, and Learn new programs, products and services through Statistics. The
range of methodologies and techniques go from simple pattern recognition and hypothesis testing, to complex multilevel generalized models, neural networks, non-parametric discriminant algorithms, and elaborated matching algorithms to ensure unbiased matched case-control studies and randomized control trials. In this session, we will be presenting a summary of the research we do and the methods we use to facilitate evidence-based informed decisions in our business.

**Monte Carlo VA Risk: Challenges of the Monte Carlo Approach in the Modeling of Variable Annuity Risk**

Edward Perry, *University of Connecticut*

Variable Annuities are complex investment guarantees sold by insurance companies to help people plan their financial lives. Given that there is no pencil-and-paper way to measure the financial risk of such a guarantee, how should we mitigate or hedge? The answer is the Monte Carlo approach. Unfortunately Monte Carlo in this context has many aspects that are very difficult to manage. Large inforce data sets, modeling the volatility skew and enforcement of correlative relationships are just a few of these challenges. In this talk we will explore ways to cope.

**Adjustment of Dependent Truncation with Inverse Probability of Weighting**

Jing Qian, *University of Massachusetts, Amherst*
Rebecca Betensky, *Harvard University*

Increasing number of clinical trials and observational studies are conducted under complex sampling involving truncation. Ignoring the issue of truncation or incorrectly assuming quasi-independence can lead to bias and incorrect results. Currently available approaches for dependently truncated data are sparse. We present an inverse probability weighting method for estimating the survival function of a failure time subject to left truncation and right censoring. Our method allows adjustment for informative truncation due to factors affecting both time-to-event and truncation. Both inverse probability of truncation weighted product-limit estimator and Cox partial likelihood estimators are developed. Simulation studies show that the proposed method performs well in finite sample. We illustrate our approach in a real data application.

**Density Estimation for Binary Sequences**

Karthik Bharath, *Ohio State University*

A histogram estimate of the Radon-Nikodym derivative of a probability measure with respect to a dominating measure is developed for binary sequences in $\{0, 1\}^N$. Necessary and sufficient conditions for the consistency of the estimate
in the mean-square sense are given. It is noted that if the dominating measure is the usual Lebesgue measure on \( \{0,1\}^N \), then the histogram estimate is not consistent.

**Attribute-dependent Partial Correlation Estimation for Gaussian Graphical Models**

Jian Guo, *Harvard University*
Xihong Lin, *Harvard University*

Gaussian graphical models explore interdependence relationships between random variables, through estimation of the corresponding partial correlation matrices. Traditionally, Gaussian graphical model assumes the partial correlations (corresponding to the edges in graphs) are homogeneous across all observations. In many scenarios, however, the edges are heterogeneous and may depend on extra factors. For example, the interactions in a gene network may be affected by environmental factors such as air pollution. In this paper, we developed a new Gaussian graphical model to incorporate the effect of extra attributes to the partial correlation and network topology. We also showed that the multi-category graphical model and time-varying graphical model in the literature can be regarded as special cases of the proposed model. The effectiveness of the proposed model was evaluated on a number of simulated examples and real examples.

**Using Medicare Claims to Study Patient-sharing Networks of Physicians**

JP Onnela, *Harvard University*
B Landon, *Harvard University*
NL Keating, *Harvard University*
M Barnett, *Harvard University*
S Paul, *Harvard University*
J O'Malley, *Harvard University*
NA Christakis, *Harvard University*

The last decade has witnessed a surging scientific interest in, and a growing public awareness of, the connectedness of modern society. Physicians too are embedded in informal networks that result from their sharing of patients, information, and behaviors. I will talk about our work on analyzing these physician-physician networks constructed from Medicare data. I will also describe how a network science based approach can be used to identify naturally occurring groups of physicians that might be best suited to becoming accountable care organizations.
Covariate Effect on Constancy Assumption in Non-inferiority Clinical Trials

Siyan Xu, Boston University
Kerry Barker, Pfizer Inc.
Sandeep Menon, Pfizer Inc.
Ralph B. D’Agostino, Boston University

Non-inferiority (NI) clinical trials are getting a lot of attention of late due to its direct application in biosimilar studies. Because of missing placebo arm, NI is an indirect approach to demonstrate efficacy of a test treatment. One of the key assumptions on the NI test is constancy assumption, that is, the effect of reference treatment is the same in current NI trials as in historical superiority trials. However, if a covariate interacts with the treatment arms, then changes in distribution of this covariate will likely result in violation of constancy assumption. In this paper, we propose four new NI methods and compare them with two existing methods to evaluate the change of background constancy assumption on the performance of these six methods. To achieve this goal, we study the impact of three elements: 1) Strength of covariate; 2) Degree of interaction between covariate and treatment and; 3) Differences in distribution of the covariate between historical and current trials have on both the type I error rate and power using three different measures of association: difference, log relative risk and log odds ratio. Based on this research, we recommend using a modified covariate-adjustment fixed margin method.

Joint Modeling of Incomplete Data with Diverse Variable Types Using Latent-Variable Models

Thomas R. Belin, University of California, Los Angeles

In incomplete data sets with many variables and diverse variable types (e.g., continuous, ordinal categorical, nominal categorical), it is challenging to develop general-purpose strategies for handling missing data. After reviewing sequential regression imputation methods (e.g., IVEWare, ICE, MICE, MIDAS) that might be viewed as competitors, this presentation will discuss joint modeling strategies based on latent-variable models that allow for the inclusion of diverse data types. In particular, we will focus on the use of models that can be fit with the help of a parameter-extended Metropolis-Hastings strategy for drawing correlation matrices in an MCMC inference framework. Illustrative examples will be presented and future directions for research in this area will be considered.

Statistical Modeling of Ambient Air Quality for IBM Smarter Planet Initiative

Youngdeok Hwang, IBM T. J. Watson Research Center
Huijing Jiang, IBM T. J. Watson Research Center
Quantitative air quality information is critical not only for policy-makers but also for ordinary residents. Ambient concentrations of pollutants are associated with various aspects, such as meteorological conditions and industrial activity. Under IBM Smarter Planet initiative, we develop a statistics-based decision making system by utilizing both physics-based computer model and various station measurements taken over space and time to mimic this complex physical system. Statistical analysis methods and ideas play a critical role in the developed system. In this talk, we describe a Smarter Planet project related to the ambient air quality modeling based on statistical spatio-temporal analysis algorithm.

**Stochastic Fixed-point Equations Arising in the Analysis of Algorithms**

Mariana Olvera-Cravioto, Columbia University

The probabilistic analysis of information ranking algorithms such as Google's PageRank naturally leads to an approximation in terms of the solution to a linear nonhomogeneous fixed-point equation on a weighted branching tree. Using an appropriate directed random graph model we show how the tree approximation is justified and therefore the analysis of linear algorithms on certain types of graphs is equivalent to the analysis of a related stochastic fixed-point equation. We also briefly mention some of the recent results that describe the asymptotic behavior of the solutions to such equations.

**Career Development Panel Discussion**

Dominique Haughton, Bentley University
Torey Vizzo, Valesta Clinical Research Solutions
Christopher Parks, Travelers
Jeffrey Young, Cigna
Naitee Ting, Boehringer-Ingelheim Pharmaceuticals Inc.

In this session, each panelist will speak for about 5 to 10 min about how to prepare CV, how to handle phone screen, and interview. After that, the session is open for Q and A.

**Bayesian Inference with Identifiable Subjects in the Prior**

Edward J. Stanek III, University of Massachusetts Amherst

Practical problems often involve estimating individual latent values based on data from a sample. We discuss an application where latent LDL cholesterol
levels of women from an HMO are of interest. We use a Bayesian model with an exchangeable prior distribution that includes subject labels, and trace how the prior distribution is updated via the data to produce the posterior distribution. The prior distribution is specified for a finite population of women in the HMO assumed to arise from a larger superpopulation. The novel aspect is accounting for the labels in the prior. We illustrate this via an example, and show how the exchangeable prior distribution can be constructed for a finite population that arose from a superpopulation. Using data that consists of the (label, response) pair for a set of women, we illustrate how conditioning on the set of labels, the sequence of labels in the sample space, and the actual response impacts the change from the prior to the posterior distributions. In particular, we show that conditioning on the actual response, alters the distribution of latent values for the women in the data, but not for the remaining women in the population.

Predictive Modeling Trends and Implications

Brian M. Stoll, Towers Watson

Predictive modeling applications in property/casualty insurance continue to grow and develop. In this session we will first explore the current state of predictive modeling, then consider some of the challenges in applying data-driven analytics to this business, company plans to overcome these challenges, and finally results from a predictive modeling case study.

Cross-Study Validation Oy Prediction Methods: a Bayesian Nonparametric Analysis

Lorenzo Trippa, Harvard University

We consider comparisons of statistical learning algorithms using multiple datasets, via leave-one-in cross-study validation: each of the algorithms is trained on one dataset; the resulting model is then validated on each remaining dataset. This poses two statistical challenges that need to be addressed simultaneously. The first is the assessment of study heterogeneity, with the aim of identifying subset of studies within which algorithm comparisons can be reliably carried out. The second is the comparison of algorithms using the ensemble of datasets. We address both problems by integrating clustering and model comparison. We formulate a Bayesian model for the array of cross-study validation statistics, which defines clusters of studies with similar properties, and provides the basis for meaningful algorithm comparison in the presence of study heterogeneity. We illustrate our approach through simulations involving studies with varying severity of systematic errors, and in the context of medical prognosis for patients diagnosed with cancer of the ovaries, using high-throughput measurements of the transcriptional activity of the tumors genes.
How to Define a Margin When the 95-95 Rule Does Not Work and How to Analyze (Sometimes You Have to Cram a Square Peg into a Round Hole)

Siyan Xu, Boston University
Kerry Barker, Pfizer Inc.
Shan Mei Liao, Pfizer Inc.
Ray Li, Pfizer Inc.

FDA guidelines have proposed to use either the 95-95 rule method or the synthesis method. However, using either of these methods requires well defined historical information based on large well controlled randomized studies. Often large well controlled historical trials are not always available. However, there is generally a wealth of information that can be used that in place of these trials. We will propose an alternative method using all historical information, including results from single arm studies. Both the 95-95 rule and the synthesis method simultaneous test for equivalence to the reference product while also testing for superiority against the putative placebo are utilized. We propose breaking this up into two different tests. One test to show test product is equivalent to a reference product. And a second test to show the strength of evidence that the test product is better than a putative placebo. This will show strength of evidence of each hypothesis individually, with less confounding from the other hypothesis. In addition, by approaching the two tests differently, the combined information gives a stronger evidence of simultaneously showing both equivalence to the reference product and superiority to the putative placebo.

Totality Test and Adaptive Design for Bio-similarity

Mark Chang, AMAG Pharmaceutical and Boston University

Biopharmaceuticals have grown tremendously in the last decade, with the global biologics market estimated to be USD 53 billion (excluding vaccines) and constituting nine percent of the global pharmaceutical market. Six of the top 10 selling drugs will expired in 2016. As the pattern cliff approaches, the door opens for the development of biosimilar products. February 2012, the FDA released two draft guidance related to biosimilars and an accompanying Q and A document to aid sponsors in the development of biosimilars. However, the guidance lacks in technical details to establish similarities in safety and efficacy. There are many similarity measures from different development stage and aspects, but there is no an agreeable global measure of biosimilarity for the decision-making. In this presentation, we proposed a totality test for biosimilarity that incorporate different similarity measures and adaptive design methods for biosimilarity trials using p-value combinations. We hope this presentation can attract more statisticians to the biosimilarity research in that direction.
Blinded Sample Size Re-Estimation Using an Adaptive Margin in Non-Inferiority Clinical Trials with a Binary Endpoint

Joe Massaro, Boston University

In active-controlled clinical trials, a possible objective is to assess non-inferiority of the experimental treatment using a pre-defined non-inferiority margin, either from a risk difference or relative-risk perspective. Sample size re-estimation (SSR) procedures protect study power by allowing sample size re-estimation based on an interim analysis using revised estimates of the design parameters such as event rate from both treatments pooled (blinded) or treatment-specific event rates (unblinded). For binary endpoints, current approaches to sample size re-estimation for non-inferiority trials focus on estimating the event rates (blinded or unblinded) at the interim and updating the sample size on the estimated event rates at the interim, without updating the non-inferiority margin, controlling alpha (Type I Error rate) at the nominal level. A procedure that adapts the absolute non-inferiority margin, and sample size based on the underlying interim observed pooled (blinded) event rate, and updates non-inferiority margin again at the final analysis based on the observed estimate of the event rate in control group at the end of the study is proposed. Our simulation results show the proposed adaptive procedures for extending a study by adding sample size, if necessary, preserves the overall type I error rate and maintains the desired power.
Abstracts of Contributed Papers

Intrusion as (Anti)social Communication: Characterization and Detection

Natallia Katenka, University of Rhode Island

A reasonable definition of intrusion is: entering a community to which one does not belong. This suggests that in a network, intrusion attempts may be detected by looking for communication that does not respect community boundaries. In this paper, we examine the utility of this concept for identifying malicious network sources. In particular, our goal is to explore whether this concept allows a core-network operator using flow data to augment signature-based systems located at network edges. We show that simple measures of communities can be defined for flow data that allow a remarkably effective level of intrusion detection simply by looking for flows that do not respect those communities. We validate our approach using labeled intrusion attempt data collected at a large number of edge networks. Our results suggest that community-based methods can offer an important additional dimension for intrusion detection systems.

Genetic Macrostructure, Replicability: What Can We Learn from the Diallel?

Alan Lenarcic, Harvard University, Jackson Labs, University of North Carolina Chapel Hill
James Crowley, Harvard University, Jackson Labs, University of North Carolina Chapel Hill
Gray Churchill, Harvard University, Jackson Labs, University of North Carolina Chapel Hill
William Valdar, Harvard University, Jackson Labs, University of North Carolina Chapel Hill

A diallel experiment is a study of the first generation of a cross of inbreds, similar to the original experiment of Mendel where he cross-bred short and tall pea strains to uncover inbred stability and hybrid vigor. Common experimental mice are fully inbred, and are thus replicable for the medical and psychological traits they model. Gene associations are often uncovered through continuous interbreeding, such as that which generated the Collaborative Cross (CC) experiment, a population of 150 strains of mice randomly derived from an original 8 scientifically valuable founders, which promises to be an important diverse resource to murine genetics. CC strains exhibit a wide array of new genetic intercombinations, though they too are replicable and subject to cross rules of the diallel. We have developed a Bayesian statistical method for the diallel to evaluate experimental results for crosses of CC strains and other model organisms.
We required an efficient but flexible algorithm, capable of quickly evaluating credibility scores, prediction metrics, and dealing with missingness and imbalance that prevent older diallel methods. With our over-specified model, we can make defensible claims on epistatic/epigenetic heritability, give model inclusion probabilities for common effects, and establish treatment*genetic interactions through a method of potential outcomes. We seek to promote the diallel as an important preliminary experiment for initiating an association study, becoming a key replicable resource for identifying sources of heritability, before larger resources should be expended on a population without a signal.

**GEE-Based Intervals for the Difference Between Two Treatment Means in Clinical Trials**

Krishna K. Saha, *Central CT State University*

This paper focuses on confidence interval construction for the difference between two treatment means in clinical trials and other similar fields. In this study, the interval methods based on the generalized estimating equation (GEE) approach and a ratio estimator approach are developed. The three other interval methods following the procedures studied for proportions are also developed. Monte Carlo simulations indicate that all the procedures have reasonably well coverage properties. However, the GEE-based interval procedure outperforms other interval procedures in terms of all three confidence interval criteria. Example in clinical trials is also presented to illustrate the proposed confidence interval procedures.

**On the Uncertainty Evaluation of the Best Linear Unbiased Predictor**

Rongheng Lin, *University of Massachusetts Amherst*
Gregory J. Matthews, *University of Massachusetts Amherst*
Andrea S. Foulkes, *University of Massachusetts Amherst*

Linear mixed effects models are widely used in the context of unevenly spaced repeated measure data and clustered data settings. When predicting cluster specific random effects is part of the inferential goal, the best linear unbiased predictor is commonly used. In this context, it is well documented that mean square errors, rather than variances should be used to evaluate the uncertainty of the best linear unbiased predictors. Our research set out to understand the relationship between mean square error and another closely related quantity, the posterior variance of the random effect given the observed data. The posterior variance given data is used as an uncertainty measure in current major mixed model package lme4 in R. We show that the posterior variance does not account for the uncertainty associated with the estimation of fixed effect during the calculation of the best linear unbiased predictor and, therefore, underestimates the prediction uncertainty. It also fails to account for correlation between prediction of clusters. Alternatively, the mean square error, which can also be expressed as
a posterior variance given the best linear unbiased predictor, accounts for this uncertainty. We derive the difference between the mean square error and the posterior variance given data and construct the correction term from current lme4 output. Theoretical development and data examples illustrate various scenarios where big gap between two approaches exists and thus the usage of the posterior variance given data should be avoided.

Investigating the Effect of Environmental Factors on Phytoplankton Count

Kalyan Das, Calcutta University
Angshuman Sarkar, Novartis India Private Limited

Several studies have discussed the effects of environmental factors on phytoplankton dynamics (Boney, 1989; De Huszar and Caraco, 1998; Kagalou et al., 1999; Hassan et al., 2004; Susanne et al., 2005; Nowrouzi et al., 2011). The influence of various factors on the seasonal appearance of phytoplankton differs significantly, with physical (such as temperature etc.) and chemical (DO, salinity etc.) factors. The main purpose of this data analysis is to study the count of Phytoplankton cell concentrations (cells/l) and its relative abundance (We have kept our main focus to explore how the count of Thalassiosira nordenskioeldii is affected by salinity and other different environmental factors by developing a suitable model. The presence of missing as well as censored data can be nicely tackled using such a general model. Analysis gives an idea how Phytoplankton cell count varies over different environments beneath the sea. The validity of such a model has been justified by a rigorous simulation study.

A Versatile Model to Analyze Life Time Data

Jie Huang, University of Maine
Ramesh Gupta, University of Maine

In life testing and survival analysis, the components are arranged in series or parallel system and the number of components is, initially unknown. Thus, the number of components is considered as random with an appropriate probability mass function. More specifically, the problem arises in cancer clinical trials where the number (N) of metastasis-competent cells (clonogens) is unknown and the event occurs as soon as one of the clonogens metastasizes. In damage models, the number of cracks is unknown and the system fails as soon as the first failure occurs. In this connection, several distributions of N have been considered in the literature, including the Poisson distribution, the Logarithmic series distribution, the COM-Poisson distribution and the power series distribution, with exponential as the baseline distribution. More recently, Gupta et.al. (2012) proposed a model with the generalized Poisson distribution as the distribution of N.

With exponential as the baseline distribution, the resulting model has de-
creasing failure rate. In this presentation, we have modeled the survival data with baseline distribution as Weibull and the distribution of N as generalized Poisson, giving rise to four parameters in the model and increasing, decreasing, bathtub and upside bathtub failure rate. For illustration, a real data example is presented. The maximum likelihood estimation of the parameters is studied and the results are compared to the existing models, especially the exponential generalized Poisson distribution of Gupta et.al. (2012).

**Homicide Rate Analysis with Dynamic Models for Near-Gaussian Random Fields Using INLA**

Renan X. Cortes, *Universidade Federal de Minas Gerais*
Thiago G. Matins, *Norwegian University of Science and Technology*
Brulio F. A. da Silva, *Universidade Federal de Minas Gerais*
Marcos O. Prates, *Universidade Federal de Minas Gerais*

Understand temporal trends of homicide rates are from great interest to better define public safety policies. However, it is not a simple task due to large variability on the response rates and/or possible temporal outlier trends. Robust time series analysis is an important issue in statistical modelling. If a dataset urges robust approach due to some specific characteristic, adopting any strategy that is not robust could produce degradation in the estimative as well as in the forecasts estimations. State-space models, also referred as Dynamic Models, is a very useful way to describe the evolution of a time series variable through a structured latent evolution system. Integrated Nested Laplace Approximation (INLA) is a recent approach proposed to perform fast Bayesian inference in Latent Gaussian Models which naturally comprises Dynamic Models. However, INLA heavily relies in the Gaussian assumption for the latent field. Recently, an extension on how to use INLA to perform Bayesian inference with some elements of the latent filed following a near-Gaussian distribution, such as Student-t, were introduced in the literature. In this paper we introduce and formalize how to use INLA to perform fast Bayesian Inference on robust (near-Gaussian) Dynamic Models. We end the paper showing the necessity of a robust dynamic analysis for the Brazilian homicide rates and interpreting the obtained results.

**Network-guided Sparse Regression Modeling for Detection of Gene by Gene Interactions**

Chen Lu, *Boston University*
Jeanne Latourelle, *Pulmonary Center, Department of Medicine and Department of Neurology, Boston University School of Medicine; The NHLBIs Framingham Heart Study*
George T. OConnor, *Boston University*
Jose Dupuis, *Boston University*
Eric D. Kolaczyk, *Boston University*
Genetic variants identified by genome-wide association studies to data explain only a small fraction of total heritability. Gene by gene interaction is one important potential source of unexplained total heritability. We propose a novel approach to detect such interactions that utilizes penalized regression and sparse estimation principles, and incorporates outside biological knowledge through a network-based penalty. We tested our new method on simulated and real data. Simulation showed that with reasonable outside biological knowledge our method performs noticeably better than stage-wise strategies (i.e., selecting main effects first, and interactions second, from those main effects selected) in finding true interactions, especially when the marginal strength of main effects is weak. We applied our method to Framingham Heart Study data on total plasma Immunoglobulin E (IgE) concentrations and found a number of interactions among different classes of HLA genes that may interact to influence the risk of developing IgE dysregulation and allergy.

**On Quinns Mathematics Is Not Science: Implications for Our Statistical Methodologies and Mathematical Biology**

Danielle Mihram, *University of Southern California*
G. Arthur Mihram, *Princeton, NJ 09542-1188*

Mathematician Quinn [AMS NOTICES 59(1): 31, 2012] concluded that mathematics is not Science, since validity in Science is external, whereas in mathematics it is internal (to the authors model). We now ask whether our mathematics is neither necessary nor sufficient for Science by examining two contemporary issues: (A) our statistical methodologies; and (B) mathematical biology. Mathematics fails to be necessary for Science: to wit, biologist C. Darwin; also, Nobel Laureate biologist KZ Lorenz: I have never in my life published a paper with a graph in it, [NATURWISSENSCHAFTEN 60: Jan 1973, p. 4]. Our mathematics is also not sufficient for Science, as was pointed out by Dean (University of Cincinnati) LT More: Mathematics deals not with Nature, but rather with abstractions [points, lines, circles: a la Euclid], [LIMITATIONS OF SCIENCE (1915): p. 151. We therefore examine our statistical methodologies (e.g., regression and statistical hypothesis-testing, to conclude that we are likely failing to meet either the criterion for Science [viz., that human activity devoted to the search for the very explanation for (i.e., for the truth about) ay particular naturally occurring phenomenon] or the conduct of the Method of our Modern Science, as delineated by mathematician Cotes in the Preface to the 1713 Cambridge edition of Newtons (quite mathematical) MATHEMATICAL PRINCIPLES OF NATURAL PHILOSOPHY: [The condition for conducting natural philosophy is] From some select phenomena they (natural philosophers) deduce, by analysis, the more simple laws follow.: i.e., observations, then mental reflexion thereon, a procedure quit often violated in the application of our regression methodology. Of course, the emphasis in applying our statistical hypothesis-testing (even if affixing a concluding P-value) seldom strives to obtain the goal of Science: viz., truth. Instead: perhaps we need to remember the
longstanding rationale for including mathematics in the secondary and tertiary educational curricula: to discipline the mind of every graduate for its use in his upcoming life as an adult.

**An Adaptive Procedure for Multiple Window Scan Statistics**

Tung-Lung Wu, *University of Connecticut*
Joseph Glaz, *University of Connecticut*

Scan statistics have been widely applied to test for unusual cluster of events in many scientific areas. It has been of practical interest on how to select the window size of a scan statistic. In this presentation, an adaptive procedure for multiple window scan statistics is introduced and the distributions are studied for independent identically distributed Bernoulli trials and uniform observations on (0,1). The idea of the procedure is to select the window sizes sequentially. An initial window size is chosen and the subsequent window sizes are then chosen depending on the value of the current scan statistic. We compare the power of our adaptive scan statistic with standard scan Statistics. Numerical results and applications for disease clusters detection are given to illustrate our procedure.
Abstracts of Poster Papers

Enhanced Archaeological Predictive Modelling in Space Archaeology: a Statistical Learning Approach on Remotely Sensed Imageries for Identifying Archaeological Sites

Li Chen, Johns Hopkins University
Carey Priebe, Johns Hopkins University
Daniel Sussman, Johns Hopkins University
Doug Comer, Cultural Site Reserve Management
James Tilton, NASA Goddard Space Flight Center

Identifying and preserving archaeological sites before they are destroyed is a very important issue. In this paper, we develop a greatly improved archaeological predictive model $APM_{enhanced}$ that predicts where archaeological sites will be found. This approach is applied to remotely-sensed multispectral bands and a single topographical band obtained from advanced remote sensing technologies such as satellites and Airborne Laser Scanning (ALS). Our $APM_{enhanced}$ is composed of band transformation, image analysis, feature extraction and classification. We evaluate our methodology on the sensor bands over Ft. Irwin, CA, USA. A nested bi-loop cross-validation and receiver operating characteristics curves are used to assess the performance of our algorithm. We first validate our method on the east swath of Ft. Irwin and then test on a separate dataset from the west swath of Ft. Irwin. A convex combination of two methodologies: $APM_{conventional}$, which has been used among archaeologists for many years, and our $APM_{enhanced}$, is demonstrated to yield superior classification performance compared to either alone at low false negative rates. We compare the performance of our methodology on different band combinations, chosen based on the archaeological importance for these sensor bands. We also compare the two types of $APMs$ in the aspects of input data, output values, practicality and transferability.

Nonparametric Bayesian Inference for the Number of Components in a Finite Mixture

Jeffrey W Miller, Brown University
Matthew T Harrison, Brown University

Dirichlet process mixtures (DPMs) are often applied when the data is assumed to come from a mixture with finitely many components, but one does not know the number of components $s$. In many such cases, one desires to make inferences about $s$, and it is common practice to use the posterior distribution on the number of components occurring so far. It turns out that this posterior is not consistent for $s$. That is, we have proven that given unlimited i.i.d. data
from a finite mixture with \( s_0 \) components, the posterior probability of \( s_0 \) does not converge to 1. Motivated by this finding, we examine an alternative approach to Bayesian nonparametric mixtures, which we refer to as a mixture of finite mixtures (MFM). In addition to being consistent for the number of components, MFMs are very natural and possess many of the attractive features of DPMs, including: efficient approximate inference (with MCMC), consistency for the density (at the optimal rate, under certain conditions), and appealing equivalent formulations (“restaurant process”, distribution on partitions, stick-breaking, and random discrete measures). Our findings suggest that MFMs are preferable to DPMs when the data comes from a finite mixture.

A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data

David S. Matteson, Cornell University
Nicholas A. James, Cornell University

Change point analysis has applications in a wide variety of fields. The general problem concerns the inference of a change in distribution for a set of time-ordered observations. For a set of multivariate observations of arbitrary dimension, we consider nonparametric estimation of both the number of change points and the positions at which they occur. We do not make any assumptions regarding the nature of the change in distribution or any distribution assumptions beyond the existence of the \( \theta \) absolute moment, for some \( (0, 2) \). Estimation is based on hierarchical clustering and we propose both divisive and agglomerative algorithms. The divisive method is shown to provide consistent estimates of both the number and location of change points under standard regularity assumptions. We compare the proposed approach with competing methods in a simulation study. Methods from cluster analysis are applied to assess performance and to allow simple comparisons of location estimates, even when the estimated number differs.

Impact of Prior Distribution Uncertainty in Multiple Imputation Inference

Valerie L. Pare, University of Connecticut
Ofer Harel, University of Connecticut

Multiple imputation is one method commonly utilized to deal with incomplete data. Imputations typically require the assignment of prior distributions to unknown model parameters. However, since there is inherent uncertainty in what the hyperparameters of these prior distributions should be, not accounting for this uncertainty will inherently lead to over-confident inferences. We propose utilizing hyperpriors to account for this uncertainty using the rules of two stage multiple imputation. In particular, we will examine the utility of this method when data is assumed to be multivariate normal.
Dynamic Compositional Modeling of Pedestrian Crash Counts on Urban Roads in Connecticut

Volodymyr Serhiyenko, University of Connecticut
John N. Ivan, University of Connecticut
Nalini Ravishanker, University of Connecticut
Md Saidul Islam, University of Connecticut

The persistent automobile dependency in USA exposes more road users, especially pedestrians, to the danger of being involved in a crash. Uncovering the temporal trend in crash counts provides a good understanding of the context for pedestrian safety. Most previous studies have used annual data to investigate the differences in pedestrian crashes between different regions or countries in a given year, and/or to look at time trends of fatal pedestrian injuries annually. Use of annual data unfortunately does not provide sufficient information on patterns in time trends or seasonal effects. This paper describes statistical methods uncovering patterns in monthly pedestrian crashes aggregated on urban roads in Connecticut from January 1995 to December 2009. We investigate the temporal behavior of injury severity levels, including fatal (K), severe injury (A), evident minor injury (B), and non-evident possible injury and property damage only (C and O), as proportions of all pedestrian crashes in each month, taking into consideration effects of time trend, seasonal variations and VMT (vehicle miles traveled). This type of dependent multivariate data is characterized by positive components which sum to one, and occurs in several applications in science and engineering. We describe a dynamic framework with vector autoregressions (VAR) for modeling and predicting compositional time series. Combining these predictions with predictions from a univariate statistical model for total crash counts will then enable us to predict pedestrian crash counts with the different injury severity levels. We also show that the dynamic models perform better than the corresponding static models. We implement the Integrated Nested Laplace Approximation (INLA) approach to enable fast Bayesian posterior computation.

Taking CO injury severity level as a baseline for the compositional analysis, we conclude that there was a noticeable shift in the proportion of pedestrian crashes from injury severity A to B, while the increase for injury severity K was extremely small over time. This shift to the less severe injury level (from A to B) suggests that the overall safety on urban roads in Connecticut is improving. In January and February, there was some increase in the proportions for levels A and B over the baseline, indicating a seasonal effect. We found evidence that an increase in VMT would result in a decrease of proportions over the baseline for all injury severity levels. Our dynamic model uncovered a decreasing trend in all pedestrian crash counts before April 2005, followed by a noticeable increase and a flattening out until the end of the fitting period. This appears to be largely due to the behavior of injury severity level A pedestrian crashes.
Fitting Large-scale GLMs with Implicit Updates

Panos Toulis, *Harvard University*
Edoardo M. Airoldi, *Harvard University*

Generalized linear models (GLMs) are among the most popular statistical models, however, fitting GLMs with thousands-to-millions of predictors raises unsolved challenges. Existing batch learning methods are not computationally feasible as they require inverting large information matrices at each iteration. On the other hand, online learning methods require careful parameter tuning and are not robust to outliers. To overcome these problems, we develop stochastic gradient descent (SGD) with *implicit updates*, which do not rely on first-order approximations and are thus provably robust to outliers. Implicit updates work at scale since their only computational overhead is solving a one-dimensional root-finding problem. Also, the proposed fitting algorithm can be applied to any GLM with a non-decreasing canonical link. Experiments on synthetic data validate our theoretical results and show the significant advantages of fitting GLMs using stochastic gradient descent with implicit updates.

Utilizing Protein Structure to Identify Non-Random Somatic Mutations

Gregory Ryslik, *Yale University*
Yuwei Cheng, *Yale University*
Kei-Hoi Cheung, *Yale University*
Yorgo Modis, *Yale University*
Hongyu Zhao, *Yale University*

Motivation:

Human cancer is caused by the accumulation of somatic mutations in tumor suppressors and oncogenes within the genome. In the case of oncogenes, recent theory suggests that there are only a few key "driver" mutations responsible for tumorigenesis. As there have been significant pharmacological successes in developing drugs that treat cancers that carry these driver mutations, several methods that rely on mutational clustering have been developed to identify them. However, these methods consider proteins as a single strand without taking their spatial structures into account. We propose a new methodology that incorporates protein tertiary structure in order to increase our power when identifying mutation clustering.

Results: We have developed a novel algorithm, iPAC (identification of Protein Amino acid Clustering), for the identification of non-random somatic mutations in proteins that takes into account the three dimensional protein structure. By using the tertiary information, we are able to detect both novel clusters in proteins that are known to exhibit mutation clustering as well as identify clusters in proteins without evidence of clustering based on existing methods. For example, by combining the data in the Protein Data Bank (PDB) and the Cat-
alogue of Somatic Mutations in Cancer, our algorithm identifies new mutational clusters in well known cancer proteins such as KRAS and PI3KCa. Further, by utilizing the tertiary structure, our algorithm also identifies clusters in EGFR, EIF2AK2, and other proteins that are not identified by current methodology.

Availability:

Handling Data with Three Types of Missing Values

Jennifer Boyko, University of Connecticut
Ofer Harel, University of Connecticut

Incomplete data is a common obstacle to data analysis in a variety of fields. Values in a data set can be missing for several different reasons including failure to answer a survey question, dropout, planned missing values, intermittent missed measurements, latent variables, and equipment malfunction. In fact, many studies will have more than just one type of missing value. Appropriately handling missing values is critical in the inference for a parameter of interest. Many methods of handling missing values fail to account for uncertainty due to the missing values which can lead to biased estimates and over-confident inferences.

One area which is still unexplored is the situation where there are three types of missing values in a study and the differences between the three types are of interest. The development of a three stage multiple imputation approach would be beneficial in analyzing studies with several types of missing values. Three stage multiple imputation would also extend the benefits of standard multiple imputation and two stage multiple imputation, namely the quantification of the variability attributable to each type of missing value and the flexibility for greater specificity regarding missing data assumptions.

Integrative Analysis of Prognosis Data on Multiple Cancer Subtypes Using Compound Group Bridge

Jin Liu, Yale University
Jian Huang, University of Iowa
Shuangge Ma, Yale University

In cancer research, profiling studies have been extensively conducted, searching for genes/SNPs associated with prognosis. Cancer is a heterogeneous disease. Examining the similarity and difference in the genetic basis of multiple subtypes of the same cancer can lead to better understanding of their connections and distinctions. Classic meta-analysis approaches analyze each subtype separately and then compare analysis results across subtypes. Integrative analysis approaches, in contrast, analyze the raw data on multiple subtypes simultaneously and can outperform meta-analysis. In this study, prognosis data on
multiple subtypes of the same cancer are analyzed. An AFT (accelerated failure time) model is adopted to describe survival. The genetic basis of multiple subtypes is described using the heterogeneity model, which allows a gene/SNP to be associated with the prognosis of some subtypes but not the others. A compound penalization approach is developed to conduct gene-level analysis and identify genes that contain important SNPs associated with prognosis. The proposed approach has an intuitive formulation and can be realized using an iterative algorithm. Asymptotic properties are rigorously established. Simulation shows that the proposed approach has satisfactory performance and outperforms meta-analysis using penalization. An NHL (non-Hodgkin lymphoma) prognosis study with SNP measurements is analyzed. Genes associated with the three major subtypes, namely DLBCL, FL, and CLL/SLL, are identified. The proposed approach identifies genes different from single-dataset analysis and has reasonable prediction performance.

**Conditional Inference for Network Analysis for Neural Spiking Data: Zero-lag Synchrony**

Dahlia Nadkarni, *Brown University*
Matthew Harrison, *Brown University*

The spiking electrical activity of simultaneously recorded neurons is often modeled as a multivariate binary time series. Inference about the fast temporal correlation structure of this multivariate time series, such zero-lag synchrony and precise lag-lead relationships, is complicated by the time varying response of the neurons to their many unobserved and correlated inputs. Classical approaches to this problem suffer from model misspecification and also incidental parameter problems, or the Neyman-Scott problem, in which the number of nuisance parameters grows with the size of the data. We develop a conditional inference approach that works well for inferring the synchrony parameters in log-linear models with non-stationary background firing rates and we develop a conditional pseudo likelihood approach that scales to larger numbers of neurons and more complex models.

**Comparing Regression Coefficients Using Incomplete Data**

Chantal Larose, *University of Connecticut*
Ofer Harel, *University of Connecticut*
Jun Yan, *University of Connecticut*

Nested regression models are used to represent two explanations of a phenomenon. Direct comparison of the regression coefficients assumes that both the reduced model and the full model are both true, which cannot be the case for the same data. An estimator was previously developed which avoids this problem but it does not address incomplete data. We extend the applicability of the estimator to incomplete datasets by performing multiple imputation.
before implementing the estimator.

**Integrative Analysis of Cancer Prognosis Data With Contrasted Group Bridge Penalization**

Xingjie Shi, *Shanghai University of Finance and Economics, Yale University*

Jin Liu, *Yale University*

Jian Huang, *University of Iowa*

Yong Zhou, *Shanghai University of Finance and Economics*

Shuangge Ma, *Yale University*

For cancer prognosis data with high-dimensional genomic measurements, integrative analysis provides a way to more effectively analyze the heterogeneous raw data from multiple independent studies and may outperform classic meta-analysis and single-dataset analysis. When marker selection is of interest, the genetic basis of multiple datasets can be described using the homogeneity model or the heterogeneity model. The heterogeneity model includes the homogeneity model as a special case and can be more flexible. Under the heterogeneity model, marker selection needs to be conducted at two levels and can be achieved using composite penalization methods, for example the group bridge. In this study, a contrasted group bridge approach is developed. The newly introduced contrast penalty can encourage the smoothness of regression coefficients for the same gene (or another functional unit) in multiple datasets. An effective iterative algorithm, which calls an inner coordinate descent iteration, is developed. Simulation shows that the proposed approach can outperform the group bridge approach with more accurate marker identification. The analysis of breast cancer and lung cancer prognosis studies with gene expression measurements shows that the proposed approach identifies genes different from those using group bridge and has better prediction performance.

**Estimating the Average Treatment Effects of Nutritional Label Use Using Subclassification with Regression Adjustment**

Michael J. Lopez, *Brown University*

Roee Gutman, *Brown University*

Propensity score methods are common for estimating a binary treatment effect when treatment assignment is not randomized. When exposure is measured on an ordinal scale (i.e., low - medium - high), however, propensity score inference requires extensions which have received limited exposure. Estimands of possible interest with an ordinal exposure are the average treatment effects between each exposure level. Using these estimands, it is possible to determine an optimal dose. Traditional methods, including dichotomization of the exposure or a series of binary propensity score comparisons across exposure pairs, are generally inadequate for identification of these effects. We combine subclassification with regression adjustment to estimate transitive, unbiased causal effects.
across an ordered exposure, and apply our method on the 2005-06 National Health and Nutrition Examination Survey to estimate the effects of nutritional label use, measured on a 5-point scale, on body mass index.

**A Bayesian Test of Independence in a Two-way Contingency Tables Under Cluster Sampling**

Dilli Bhatta, *Worcester Polytechnic Institute*

A Bayesian method is implemented to construct a test of independence of two variables in a two-way contingency table which is obtained from a two-stage cluster sampling design. The association between the two categorical variables is studied (a) without covariates and (b) with covariates at both unit and cluster levels. The key idea of our Bayesian test of independence is to convert the cluster sample into an equivalent simple random sample which provides a surrogate of the original sample. Then a simple formula for the Bayes factor is used to make inference about independence. To construct the test, hierarchical Bayesian models (multinomial and multivariate logit) are developed to convert the complex survey data to the surrogates (simple random samples). Efficient Markov chain Monte Carlo algorithms (e.g., Gibbs samplers) are constructed to fit the models. The methods in (a) are applied to the Third International Mathematics and Science Study to assess the association between the mathematics/science test scores and communities for third grade U.S. students and in (b) to the Trend in International Mathematics and Science Study to assess the association between the mathematics and science scores for fourth grade U.S. students by community. The goodness of fit of these models and extensive simulation studies show that our approach is reasonable.

**Population Based Convergence Criterion for Self-Organizing Maps**

Benjamin Ott, *University of Rhode Island*
Gregory Breard, *University of Rhode Island*
Lutz Hamel, *University of Rhode Island*

Self-organizing maps (SOMs) are a type of artificial neural network extensively used as a data mining and analysis tool in a broad variety of fields including bioinformatics, financial analysis, signal processing, and experimental physics. They are attractive because they provide a simple yet effective algorithm for data clustering and visualization via unsupervised learning. A fundamental question regarding self-organizing maps is the question of convergence, or how well the map models the data distribution after training. Here we introduce a population based convergence criterion: the neurons of the map represent one population and the training data represents another population. The map is said to be converged if the neuron population and the training data population appear to be drawn from the same probability distribution. This can easily be tested with standard two-sample tests. Our poster aims to present this
approach and to demonstrate applications of this new convergence criterion to non-trivial data sets. The proposed convergence criterions advantage over Cottrell et als statistical tools for assessing SOM stability and reliability is that the proposed convergence criterion is much less computationally expensive than Cottrells criterion. Furthermore, preliminary investigations have shown that the proposed criterion is also more conservative than Cottrells criterion when used to determine when a SOM has been sufficiently organized. The advantage over Bishops generative topographic mapping (GTM), Verbeeks generative self-organizing map (GSOM), and approaches minimizing energy function which have been imposed on the SOM is that the population based convergence criterion does not modify the SOM algorithm, is demonstrably effective, and is much simpler, seemingly in line with the SOM algorithm which itself is very straightforward. The poster also aims to demonstrate that our convergence criterion can be considered a model selection criterion by demonstrating that SOMs which achieve a higher level of convergence as indicated by our convergence criterion better model the probability density of the training data and, in general, are better maps which exhibit more well-defined clusters. In addition, the poster provides a high level tutorial on the usage of an R package currently under development at the University of Rhode Island which employs these techniques. With our new convergence criterion and our novel starburst visualization technique, the SOM may be viewed as a statistically sound data and cluster visualization system similar to hierarchical clustering. The clusters we can observe with starbursts are visualizations of structures that appear in the training population.

General Random Utility Models for Social AndPersonalized Choice

Hossein Azari Soufiani, Harvard University
David C. Parkes, Harvard University
Lirong Xia, Harvard University

Random utility theory models an agents preferences on alternatives by drawing a real-valued score on each alternative (typically independently) from a parameterized distribution, and then ranking the alternatives according to scores. A special case that has received significant attention is the Plackett-Luce model, for which fast inference methods for maximum likelihood estimators are available. We develop conditions on random utility models (RUMs) that enable inference within a framework through MC-EM, providing concave log-likelihood functions and bounded sets of global maxima solutions. We also extend the theory and the algorithm to generalized random utility models (GRUMs) to take into account agents and alternatives features, and point out an application of preference elicitation using ideas from Bayesian experimental design. Results on both real-world and simulated data provide support for the scalability of the approach and capability for model selection among random utility models including Plackett-Luce.
Bayesian Predictive Model Assessment and Selection with Bregman Divergence

Gyuhyeong Goh, University of Connecticut
Dipak K. Dey, University of Connecticut

In the Bayesian model determination, the predictive distribution or density have been considered as the main tools. In this paper, we introduce a new criterion called the Bregman divergence criterion, which provides a general Bayesian predictive model selection or assessment measure. Using Monte Carlo approach, we develop an efficient estimator of (conditional) predictive distribution and density in the Bayesian framework. To check calibration, a generalized probability integral transform is proposed. The tool can be applied to independent data as well as dependent data. Some illustrative examples are provided via a linear regression, a logistic regression, and a longitudinal data model.

A Structured Covariance Model for Quantifying Transcriptional Control of Protein Levels

Alexander Franks, Harvard University
Gabor Csardi, Harvard University
D. Allan Drummond, University of Chicago
Edoardo Airoldi, Harvard University

In biology, researchers often assess the strength of linear relationships between biological variables by computing their empirical correlation. However, without a statistical model that incorporates features of the data, these estimates are often very biased. We show that studies which do not take into account measurement noise, correlated error or missing data lead to attenuated estimates of correlation. To address these issues, we develop a method for estimating the covariance matrix of measured quantities with structured errors in the presence of non-ignorable missing data. We demonstrate our method in a case study on the role of transcriptional control of protein production in the yeast, S. cerevisiae. We find that the correlation between mRNA levels and protein abundance is high, suggesting that post-transcriptional regulation is less prevalent than previously reported.

Approximating the Convolution of Lognormal Random Variables within a MCMC Sampler

Guillaume Basse, Harvard University
Eric Solis, Harvard University
Edoardo Airoldi, Harvard University

In this paper, we seek an approximation for the convolution of lognormal random variables for use inside of an MCMC sampler. We show empirically that
approximating the convolution by a lognormal random variable is reasonable under certain conditions, and we give simple rules summarizing these conditions. We then improve on an existing method by introducing Laplace approximations, and show using extensive simulations, that our method performs better than its main competitors under the rules mentioned above.

Bayesian Degree-corrected Stochastic Block Models for Community Detection

Lijun Peng, Boston University
Luis E. Carvalho, Boston University

We discuss a degree-corrected version of a stochastic block model that aims to achieve a better resolution for the problem of community detection in networks. Our model is similar to Newman & Karrer’s but we model network adjacencies using a logistic regression on community labels and nodes effects. We follow a Bayesian approach, adopt a data augmentation strategy with latent Polya-Gamma variables and conduct inference based on a centroid estimator that formally addresses label identifiability issues. We demonstrate the novel proposed model and estimation on real-world as well as simulated network datasets and show that the centroid estimator is more flexible, principled, and yields smaller misclassification rates when compared to the maximum a posteriori estimator and Binder’s estimator. Finally, we offer a few concluding remarks on the model implementation and directions for future work.

Incorporating Spatial Dependence in Regional Frequency Analysis

Zhuo Wang, University of Connecticut
Jun Yan, University of Connecticut
Xuebin Zhang, Environment Canada

The efficiency of regional frequency analysis (RFA) is undermined by the spatial dependence among the sites under consideration, which is usually ignored in parameter estimation. We propose to incorporate spatial dependence in the estimation procedure of RFA by pairwise bivariate extreme value distributions. Estimators of model parameters and return levels from a pairwise likelihood approach is more efficient than those from traditional RFA. In a large scale simulation study, we compared the efficiency of three RFA methods: L-moments, independence likelihood, and pairwise likelihood. Various spatial dependence model and dependence levels were considered. The pairwise likelihood approach was found to give estimators with the smallest mean squared error if the dependence model was correctly specified. Even when the dependence model was misspecified, the pairwise likelihood approach was still competitive in relative to the other two methods. The method is applied to a RFA with data from Southern Ontario.
On the Characterization of a Class Of Fisher-Consistent Loss Functions and Its Application to Boosting for Hierarchical Outcomes

Matey Neykov, Harvard University
Tianxi Cai, Harvard University

Fisher-Consistent Loss functions play an important role in Decision Making Theory. In this talk we focus on a characterization of a broad class of Fisher-Consistent Loss functions. Following closely Zou et al (New Multicategory Boosting Algorithms Based on Multicategory Fisher-Consistent Losses), we generalize existing results on the class of loss functions that achieve Fisher consistency. We also propose a generic iterative procedure that converges to the minimizer of the loss function and illustrate the new proposal with a generalized boosting algorithm. We further extend our proposal to enable efficient classification of hierarchically structured multiple outcomes with a hierarchical boosting algorithm. We demonstrate that the new proposal, leveraging the tree structure of the outcomes, attains optimal classification accuracy with respect to a class of pre-specified cost parameters.

An EM Algorithm for the Matrix Normal Distribution

Hunter Glanz, Boston University
Luís Carvalho, Boston University

As the size and dimensionality of data sets grow, increasingly sophisticated techniques for estimation and analysis help ensure more complete exploitation of the information in the data. To this end, we should take advantage of structure in the data whenever possible in order better characterize this information. In this article we consider a random vector $X$ distributed normally with mean $\mu$ and covariance $\Sigma = \Sigma_c \otimes \Sigma_s$. This type of covariance structure, which is characterized by the matrix normal distribution, is common in spatio-temporal data analysis and can be used when there exists some sort of separability in the dimensions of the data. We review the previously established maximum likelihood estimates for these parameters, and then consider the situation involving missing data. An expectation-maximization algorithm is derived for estimating these parameters and tested on simulated data as well as remotely sensed data from MODIS.

Estimating Network Degree Distributions Under Sampling: an Inverse Problem, with Applications to Monitoring Social Media Networks

Yaonan Zhang, Boston University
Eric Kolaczyk, Boston University
Bruce Spencer, Northwestern University

Networks are a popular tool for representing elements in a system and their
interconnectedness. Many observed networks can be viewed as only samples of some true underlying network. Such is frequently the case, for example, in the monitoring and study of massive, online social networks. We study the problem of how to estimate the degree distribution—an object of fundamental interest—of a true underlying network from its sampled network. In particular, we show that this problem can be formulated as an inverse problem. Playing a key role in this formulation is a matrix relating the expectation of our sampled degree distribution to the true underlying degree distribution. Under many network sampling designs, this matrix can be defined entirely in terms of the design and is found to be ill-conditioned. As a result, our inverse problem frequently is ill-posed. Accordingly, we offer a constrained, penalized weighted least-squares approach to solving this problem. A Monte Carlo variant of Stein’s unbiased risk estimation (SURE) is used to select the penalization parameter. We explore the behavior of our resulting estimator of network degree distribution in simulation, using a variety of combinations of network models and sampling regimes. In addition, we demonstrate the ability of our method to accurately reconstruct the degree distributions of various sub-communities within online social networks corresponding to Friendster, Orkut, and LiveJournal. Overall, our results show that the true degree distributions from both homogeneous and inhomogeneous networks can be recovered with substantially greater accuracy than reflected in the empirical degree distribution resulting from the original sampling.

Polynomial Parameterization of Single Nucleotide Polymorphism in Genetic Association Studies of Quantitative Traits

Harold Bae, Boston University
Paola Sebastiani, Boston University

In a genome-wide association study (GWAS), the association between each single nucleotide polymorphism (SNP) and a quantitative trait is tested using linear regression under a specific genetic model, which can assume 2-df general, dominant, recessive, co-dominant, or additive mode of inheritance of the causal allele. When the inheritance pattern is not known, selecting the correct genetic model for each SNP requires fitting five models and choosing the model that describes the data best. However, this approach accompanies computational burden with the genome-wide data, and the optimal method for choosing the best model is not clear. The common practice, which is testing the additive genetic model alone, may fail to capture some true associations if the mode of inheritance is not additive. Therefore, we propose a polynomial parameterization, where we include the first and second degree polynomials of the additive coding of the SNP genotype (0, 1, or 2) in the model, in a Bayesian framework. We show that there is a mathematical relationship between the polynomial model and each of the five possible genetic models. There is a one-to-one transformation between the polynomial and 2-df general model, while the transformation between the polynomial and other four models is constrained by a linear contrast
of the parameters in the polynomial model. The model selection is performed by computing the marginal likelihood of the five models, which we derive using the relationship between the polynomial and other genetic models. Given the model with the highest marginal likelihood or Bayes factor, we derive the estimates of genetic effects from the polynomial model. Finally, we present the results of 10,000 simulations with 841 unrelated individuals and 50,000 SNPs with varying minor allele frequencies. The proposed method provides a direct way of comparing different genetic models, while reducing the amount of computations.

F-tests with Incomplete Data Under the Multiple Regression Setup

Ashok Chaurasia, University of Connecticut
Ofer Harel, University of Connecticut

Tests for regression coefficients such as the partial F-tests are common in applied research. When dealing with incomplete data, the task of conducting F-tests remains elusive. In this paper we propose a method based on the coefficient of determination to perform partial F-tests with multiply imputed data. Our proposed method can be applied for conducting the "global" F-test (test for all regression coefficients equal to zero), partial F-test (for one or more coefficients, but not all, equal zero), or for equality of regression coefficients. The proposed method is evaluated using simulated data and applied to suicide prevention data.

Cause & Effect: Standardizing Knowledge through Statistics

Mike Sagherian, University of Connecticut
Tania B. Huedo-Medina, University of Connecticut

Meta-analysis has become increasingly popular over the past 30 years, and is now widely used to review data gathered across nearly all scientific domains. Meta-analytic procedures pool the results of independent studies on a given research topic, pursuing three main objectives: (a) synthesizing different studies effect size (ES) values to obtain a weighted mean, (b) assessing the consistency of the ES values, and (c) in the case of inconsistency, using moderator variables to explain the variability in the ES values. Because studies often examine the same phenomenon using different measures and because many study reports include only inferential statistics, meta-analysts often resort to standardized ES metrics. Even when every study uses the same measure, meta-analysts often use these transformations to enable comparisons across phenomena on the same metric. When the phenomenon concerns a comparison of two means, meta-analysts routinely calculate the ES as a standardized mean difference, $d$, with an estimator denoted as $d$. This estimator is used to measure the degree of change between repeated measures or the difference between two groups, with the assumption that the measures follow a normal distribution. In its between-groups form, $d$
can be calculated from any two groups whether they are experimental or not; it is assumed that the individuals in the compared groups are independent. In its repeated-measures or within-subjects form, $d$ assumes that the observations are dependent, and while some extant meta-analytic procedures account for this dependency, many others do not; scholars will often integrate both types of estimates in a single meta-analysis. Some other studies report the data using other statistics like percentages, odd ratios to analyze the same ultimate relationship between variables. The numerous methods of calculating the ES and their variances are known to vary. Therefore, a guide to calculate a common metric based on the different study design and statistical information that the study reports will facilitate the work of the meta-analyst or other researchers that may need to calculate ES from other studies using secondary aggregated data. The purpose of this work is to present an ES calculator and its guide that facilitate the calculation of ESs for individual studies using aggregated data. The step-by-step guide will insist in using the calculator to convert various study statistics to a standardized ES metric, thus creating the ability to compare outcomes reported as different statistical measures. There are three main study designs that the calculator incorporates: two groups with pretest and posttest data, two groups with posttest data only, and one group with pretest and posttest data. The main statistics that can be converted through the calculator are Means & SDs, Means & SEs, Means & CIs, F-tests, Chi-square tests, t-tests, Z-tests, Betas/correlations, Proportions, and Odds Ratios. By standardizing statistics and allowing for their comparability, one can analyze and aggregate outcomes within an individual study and also across an entire range of studies.

Local Algorithm to Estimate Graphons by Stochastic Blockmodels Approximation

Thiago Costa, Harvard University
Edoardo Airoldi, Harvard University

This paper introduces a non-parametric method of estimating graphons based on stochastic blockmodel approximation. For any symmetric measurable $w : \mathbb{R}^2 \rightarrow [0, 1]$, one can define a random graph model $G(n, w)$ with $n$ vertices by first sampling labels $u_i \sim \text{Uniform}[0, 1]$ to each vertex $i \in \{1, \ldots, n\}$, then connecting any two vertices $i$ and $j$ with probability $w(u_i, u_j)$. We consider the problem of recovering $w$ from a set of observations of $G(n, w)$. Our algorithm finds a step function $w' : \mathbb{R}^2 \rightarrow [0, 1]$ that consistently approximates $w$ as $n \rightarrow \infty$. The theory of graph limits guarantees that, if our estimation is good enough, the resulting stochastic blockmodel preserves many properties of the observed graphs, such as the density of subgraphs. The procedure is fast and applicable to very large graphs. The idea is based on locally computing a similarity score between vertices, then clustering similar vertices to estimate the blocks of $w'$. We test the method in different types of graphs, such as small world and scale free, as well as in networks generated from stochastic blockmodel and latent space model, and we find that our algorithm outperforms existing
models. We also present, as an application of the proposed model, a fast method to count subgraphs in the observed networks.

**A Generalized Robust Control Problem and Stochastic Differential Utility**

Jingshu Liu, *Boston University*
Marcel Rindisbacher, *Boston University*

In this paper, we establish the equivalence between a generalized robust control problem with the stochastic differential utility problem. This new robust control problem generalizes the distance measure from the relative entropy to the alpha family of divergence measure. It has an integrated utility form, allowing for rich interactions between utility from consumption and loss from measure divergence. This problem nests the problem studied in Skiadas (2003) where the ambiguity averse agent balances between the sum of consumption utility and relative entropy. This generalized robust control problem is equivalent to a stochastic differential utility problem, with closed form optimal solutions available for an agent with CRRA utility. We allow for stochastic investment opportunities, with inter-temporal dynamic hedging components derived through a Malliavin calculus approach. We numerically evaluate the optimal consumption and dynamic hedging components through algorithms for forward-backward stochastic differential equations. Numerical results show that the agent’s attitude towards ambiguity has significant effects on optimal consumption and portfolio allocation, and her attitude towards ambiguity can be connected with the preference over the timing of uncertainty resolution. We explore the asset pricing implication in a pure exchange market where the representative agent solves the generalized robust control problem. A loving attitude towards ambiguity can help explain the equity premium puzzle.

**Inference Based on Estimating Equations for Logarithmic Autoregressive Conditional Duration (ACD) Models**

James Anderson, *University of Connecticut*
Lilian Cheung, *University of Connecticut*

There is considerable recent interest in statistical analysis of durations or times between consecutive events of interest, \( x_i = t_i - t_{i-1}, i = 1, 2, \ldots \), with \( t_i \) representing the time of the \( i \)th event and \( t_0 \) being the starting time. There are interesting applications in several disciplines; for instance, examples in finance include estimation of patterns in durations until a stock returns to a benchmark price, or in times between similar price movements of a stock. Several models have been discussed in the literature for the analysis of the positive-valued durations from high frequency financial data (Pacurar, 2006). The basic ACD model proposed by Engle and Russell (1998) uses a linear parameterization of the expectation of \( x_i \) conditional on past information, denoted by \( F_{i-1}^x \). This re-
search investigates inference for two types of log ACD models via the estimating equations approach described in Thavaneswaran et al. (2012). This approach only requires specification of the first four moments of the error distribution, but not its PDF. We use R code to compute recursive estimates of parameters from low-order models under different assumptions on the errors. We compare models using predictive accuracy. This work is based on undergraduate research mentored by Prof. Nalini Ravishanker, University of Connecticut.

Development of Healthy Eating Preference Index: Evidence of Content and Predictive Validity

Mastaneh Sharafi, University of Connecticut
Valerie B Duffy, University of Connecticut
Robin J Miller, University of Connecticut
Suzy B Winchester, Brown Center for Study of Children at Risk, Women & Infants Hospital
Mary C Sullivan, University of Rhode Island
Tania B Huedo-Medina, University of Connecticut

Background: Food groups and nutrients have been shown to associate with chronic diseases in several studies. However, in reality we do not consume food and nutrients separately. To reflect overall dietary patterns, a number of quality indices were developed using food frequency questionnaire or dietary recalls. A well-know example is Healthy Eating Index (HEI) that is developed by USDA and has been shown to be related with health outcomes. This study aimed to develop diet quality index using food preference survey, which is a fast screener of habitual eating and assess its validity. METHOD: The sample size was 168 23-y-old adults born preterm or full term. Subjects reported their preference for food and non-food items using the hedonic general Labeled Magnitude Scale (gLMS). The preference scores ranged from -100 to +100. Following the same concept as HEI study, different groups of food were created based on the literature and a latent class approach. HEPI was formed assigning weights to the following food groups/items of preference survey: fruits and vegetables (+3), protein (+1), fat (-3), sweets (-3), salty (-3), wine (+2), and variety (+3). The weights were chosen based on their importance for health as recommended by dietary guideline 2010. Taking an average of all weighted scores, HEPI score was formed with higher score indicating greater preference toward healthy eating or better diet quality. Adiposity, blood pressure and blood lipids were measured to assess risk factors for chronic disease and were used to test the predictive validity. RESULTS: The preference score obtained with HEPI was normally distributed with Mean = -36 and SD = 37. Total Chronbachs alpha for HEPI was 0.63. Alpha, for all the food groups was greater than 0.7 except for FV (=0.56) and salty food (= 0.67). Latent variable modeling showed that those who had higher HEPI score had higher level of HDL and lower levels of systolic blood pressure and Total cholesterol/HDL. For adults born preterm, HEPI was lower than control. HEPI also mediated the association between term status and
systolic blood pressure. CONCLUSIONS: Overall, the present study showed good validity of HEPI by confirming the expected associations between diet quality and some of the main risk factors for chronic disease.

**Defect Growth Prediction in Rolling Bearings Based on Approximate Entropy**

Peng Wang, *University of Connecticut*
Robert X. Gao, *University of Connecticut*

Approximate Entropy (ApEn), as a statistic parameter, could effectively measure complexity and regularity of a time series in a multiple dimension space by measuring similarity of reconstructed vectors. And this characteristic makes ApEn an effective approach for characterizing the structural degradation of bearing, and severity of the defect. So this paper is mainly focused on the analysis of vibration signal measured on rolling bearing based on Approximate Entropy. First of all, a qualitative empirical formula is proposed to investigate the relationship between ApEn and different failure patterns and defect severity, where failure patterns are recognized by characteristic frequency and defect severity is represented by the ratio of signal to noise (SNR). It is shown that ApEn values would decrease with the degradation of defect under certain failure pattern. And for different failure patterns under certain SNR, ApEn values would decrease with the increase of characteristic frequency. Then three steps of ApEn variation with defect degradation from initial failure period to final destruction is proposed, and a life cycle experiment is introduced to evaluate ApEn as an attractive parameter for remaining service life prognosis of rolling bearing. Simulation and experimental analysis results show that ApEn is effective for bearing defect diagnosis and prediction.

**A Noise-assisted Data Processing Method in Weak Signal Detection: Stochastic Resonance and Its Adaptive Scheme**

Rui Zhao, *University of Connecticut*
Robert Gao, *University of Connecticut*

Noise is generally referred to as something undesirable. For example, in signal detection, noise is considered as a disturbance that reduces the information transmission. The traditional signal processing methods are all aimed at filtering or even masking noise to extract the weak signal. However, useful information may be influenced or destroyed with the weakening of noise. On the other hand, in recent years, researchers have investigated some noise-assisted data processing methods to enhance the detectability of the original signal with Stochastic Resonance (SR) as a typical representative. In this paper, a new adaptive SR algorithm is presented, which can obtain the frequency content of a periodic component submerged in a noisy signal without prior knowledge of the component. The proposed new adaptive scheme successfully solves the
problem about selecting system parameters to produce SR adaptively. And the introduction of the cascading module improves the precision for measuring the frequency value of a weak component. Simulations and experiments have confirmed that the proposed method can extract the true signals submerged in externally imposed, undesirable noise with high precision and robustness.

**Spatially Weighting Genes in Variable Selection for Genome-Wide Association Studies**

Ian Johnston, *Boston University*
Luis E. Carvalho, *Boston University*

Motivated by the important problem of detecting association to traits in genome-wide association studies (GWAS), we present a novel Bayesian model that establishes a hierarchy between single nucleotide polymorphisms (SNPs) and genes by defining weights according to gene lengths and distances from genes to markers. The proposed hierarchical model enables the selection of both SNPs and genes that are most likely to be associated with a given trait. To this end we formally conduct posterior inference in the form of variable selection. We describe the relationship between our model and other popular penalized regression models such as lasso and ridge and then compare the performances of each model on simulated GWAS data. We conclude with a discussion on preliminary results on a dataset generated using real SNP data from the Wellcome Trust Case Control Consortium.

**The Network You Keep: Graphlet-Based Discrimination of Persons of Interest**

Saber Shokat Fadaee, *Northeastern University*
Nikos Passas, *Northeastern University*
Ravi Sundaram, *Northeastern University*

The past decade has seen a dramatic growth in the popularity and success of social networks. Technological advances have made it possible to capture a digital trail of the interactions and linkages between individuals. The question of how individual behaviors and interactions lead to the structure and evolution of social networks has been the focus of intense research investigations for a while now. In this work we take a fresh perspective and address the problem from the opposite end - namely, what can we say about individuals based on knowledge of their network of interactions and connections. And, instead of studying regular individuals and popular social networks (such as Twitter, Facebook etc..) We focus on a new dataset, networks of ”persons of interest”. This is a dataset of interest to law-enforcement agencies, financial institutions, etc. It is particularly relevant to the problem at hand because, it is often easier to infer the network of connections from the public Internet rather than the specific activities of individuals. We invested significant effort in creating a comprehensive
dataset by gathering and fusing information from a variety of public and commercial sources, including (but not restricted to): UN, World-Check(Thomson Reuters), Interpol, Factiva(Dow-Jones), OFAC, police websites, etc. Our final dataset was comprised of seven hundred thousand persons of interest with four million connections among them. We then filtered this dataset down to slightly less than six hundred thousand individuals who fell into one of the following 5 categories: 1. Suspicious Individuals (appeared on sanctioned lists, arrested, detained etc., but not convicted) 2. Convicted Individuals 3. Legal Individuals (lawyers, attorneys, judges, professionals who have been charged or debarred) 4. Political Individuals (elected officials, heads of parties, holders of political positions now or in the past) 5. Terrorists With the filtered dataset in hand we proceeded to address the specific question of inferring the category given an induced subgraph drawn from one of the 5 categories. Our main discovery is that a variant of Graphlet Decomposition gives us a statistical test that discriminates between the categories with high power. Graphlet Decomposition is a relatively new technique that we adapt to our setting using graph powers. The resulting technique is high both in discriminating power as well as efficiency. As a point of comparison we created a similar discriminant based on the traditional Eigenvalue Decomposition and found that it has low power to discriminate between categories and also takes a long time to run. To be specific, if we used the two techniques to recover the categories from the raw dataset using clustering we found that the dimensions of the Graphlet-basis was one-fortieth the dimensions of the Eigen-basis while generating clusters with intra-cluster variance a fourth that of the Eigenvalue Decomposition. Further, the Graphlet-basis has a natural representation in terms of cliques with implications for the cell structure of the individuals of interest. Our finding indicate the (high) potential for the development of statistics-based technologies for categorizing and drawing inferences purely from network structure, about individuals (even those whose activities are not self-disclosed on the public Internet).

Behavioral Versus Biological Measures on Sexual Risk: a Systematic Review and a Comparison of Intervention Effect Sizes

Samantha Russo, University of Connecticut
Ofer Harel, University of Connecticut
Sagherian, M., University of Connecticut
Tania B. Huedo-Medina, University of Connecticut

Background: Sexual risk behavior measures can be biological using laboratory assays or behavioral using self-report. An accurate assessment of sexual risk behavior is necessary to make a correct evaluation of sexually transmitted behavioral (STIs) interventions. Many studies in the literature at hand use self-reports of sexual behavior to evaluate sexually transmitted infections without any other measures. The main goal of this work is to systematically review the different measures that have been used in the HIV prevention interventions literature and compare biological and behavioral outcomes of behavioral
interventions to increase condom use and reduce sexually transmitted infections (STIs), including HIV. Methods: A systematic search was conducted using electronic databases and the citations section of other relevant published papers. An application of the differences between effect sizes from behavioral biological measures was conducted using the datasets from two comprehensive published meta-analyses that were supplemented by additional Diffusion of Effective Behavioral Interventions (DEBIs). The studies were included if they examined specific measures of interest. Studies were retrieved from electronic databases using the reference section of the two meta-analyses previously published and the Centers for Disease Control and Prevention (CDC). Fifty-seven studies with 276 separate interventions were included. Independent raters calculated effect sizes. Weighted effect sizes for both measures, STIs, including and without including HIV, and for condom use were calculated under random-effects assumptions. Other intervention (e.g., length of the intervention) and sample characteristics (e.g., mean age) were coded following a prior coding form that was developed and tested. Finally, specificity and sensitivity of the behavioral measure, condom use, were obtained. Positive effect sizes indicated increased condom use and fewer incident STIs, including HIV. RESULTS.

Conditional Inference for Edge Covariate Effects in Sparse Networks

Daniel Klein, Brown University
Matt Harrison, Brown University

Many systems in biology (connectomes), ecology (food webs), sociology (social networks), and other applied sciences can be modeled as random directed networks. A simple class of models is defined by taking the edges to be independent Bernoulli distributed with their log-odds a linear function of observed edge-level covariates and a global offset. The often extreme degree heterogeneity observed in real-world networks has prompted the inclusion of degree-correction terms (Karrer and Newman, 2010), giving edge probabilities between node $i$ and node $j$ as

$$\logit P^{(N)}_{ij} = \alpha_i + \beta_j + \kappa^{(N)} + \theta^T x_{ij},$$

a logistic regression model that extends the Rasch model from psychometrics. Questions of consistency can be posed in this setting by letting the number of nodes $N$ grow large and letting $\kappa^{(N)}$ vary to give a sequence of networks with the desired scaling. With dense scaling (expected degree linear in $N$), MLE inference for the parameter of interest $\theta$ is well-behaved. However, with sparse scaling (expected degree constant), MLE inference fails due to the presence of nuisance parameters $\alpha$ and $\beta$.

We propose using conditional inference for this problem. Conditioning on the observed out- and in-degree sequences eliminates $\alpha$, $\beta$, and $\kappa$ from the model. Computing the exact conditional likelihood is intractable but the proposal distribution of a sequential importance sampler (Harrison and Miller, preprint) serves as an excellent surrogate likelihood surface. Further, the importance
sampler can be used to produce confidence intervals whose validity is robust to the quality of the proposal distribution (Harrison, 2012).

The failure of MLE inference and success of conditional inference in this setting are demonstrated on synthetic data for which the stated model is correct. Conditional inference also performs well under mild model misspecification, e.g., networks sampled with fixed degree sequence. The proposed approach is then applied to a biological data set, the chemical synapse connectome of the nematode *C. elegans* (279 nodes, 2194 edges), for which several meaningful covariates (cell body location, lineage distance, and sensory/motor/interneuron pair class membership) are available. This analysis demonstrates that conditional inference for networks scales to realistically sized problems.
The 2013 New England Statistical Symposium would not be possible without the generous support of our sponsors. Please join us in thanking the following organizations for their contributions to this year’s symposium.

The first New England Statistics Symposium was held at the University of Connecticut. We have been continuing the tradition of hosting the Symposium on alternate years while inviting other institutions throughout New England to host it in years in between.