The 26^{th} New England Statistics Symposium

Department of Biostatistics
Department of Mathematics and Statistics
Boston University

April 20-21, 2012

Welcome!

The Department of Biostatistics and the Department of Mathematics and Statistics are proud to welcome you to Boston University for the 26^{th} New England Statistics Symposium. We are thrilled to be jointly hosting this annual event this year and hope that it proves to be an intellectually stimulating and enjoyable experience for all.

The program consists of two plenary speakers, and two sets of parallel sessions, both invited and contributed. In addition, we are holding both the usual graduate student paper competition and, new this year, an undergraduate student poster competition. Finally, we are pleased to offer three short courses. Details of the program may be found on the following pages (also available at http://math.bu.edu/ness12/program.html).

We would like to thank all who have helped to make NESS 2012 a reality this year. In particular, we would like to thank the following industry sponsors for their generous support of the symposium and the student paper and poster competitions:

- Harvard Clinical Research Institute (HCRI)
- PROMETRIKA
- SAS Institute Inc., JMP Division

In addition, we would like to thank the College of Arts and Sciences and the Departments of Biostatistics and of Mathematics and Statistics, at Boston University, and the Boston Chapter of the American Statistical Association, for their kind support.

Welcome to NESS 2012 at Boston University and enjoy!

Josée Dupuis
Eric Kolaczyk
Co-Chairs

Schedule, Friday, April 20, 2012

1:00-5:00pm: **Short Course 1**

Location: Crosstown Center Room 460A

Topic: Adaptive Clinical Trial Design and Simulation

1:00-5:00pm **Short Course 2**

Location: Crosstown Center Room 460

Topic: Statistical Analysis of Network Data

1:00-5:00pm **Short Course 3**

Location: Crosstown Center Room 462

Topic: Stochastic Modeling of Limit Order Books: a Journey Across Time Scales

Schedule, Saturday, April 21, 2012

09:15am **Registration & Coffee**

Location: Lobby, BUSM Instructional Building L, 72 E. Concord Street

09:45 - 10:00am **Welcome and Opening Remarks**

Location: Bakst Auditorium, BUSM Instructional Building L

10:00 – 11:00am **Keynote Presentation: Rick Durrett, Duke University**
Title: Branching Process Models of Cancer

Location: Bakst Auditorium, BUSM Instructional Building L

11:00am **Coffee Break**

11:30am **Parallel Paper Sessions AM**

Invited Sessions

A1 Topic: Recent Advances in Multivariate and Functional Data Analysis

Location: L203

A2 Topic: Teaching Business Statistics: Best Wisdom and Biggest Challenges

Location: L201

A3 Topic: Statistics and Clinical Trials

Location: L210

A4 Topic: Sampling and Inference from Ignorable and Non-Ignorable Network Sampling Designs

Location: L206

A5 Topic: Machine Learning

Location: L209

Contributed Sessions

A6 Topic: Network Analysis and Covariance Matrix Estimate

Location: L214

A7 Topic: Applied Statistics

Location: L212

Student Paper Competition Sessions

A8 Topic: Statistical Genetics and Bioinformatics

Location: L211

A9 Topic: Applied Probability

Location: L213

1:00 – 2:30pm **Lunch**

Undergraduate Student Poster Competition Session

Location: L-hallway, 2nd Floor

2:30 – 3:30pm **Keynote Presentation: Robert Kass, Carnegie Mellon University**
Title: The Central Role of Modern Regression in Statistical Thinking about Neural Spike Trains

Location: Bakst Auditorium, BUSM Instructional Building L

3:30 – 4:00pm **Coffee Break**

4:00 – 5:30pm **Parallel Paper Sessions PM**

Invited Sessions

B1 Topic: Operation Research and Statistics: At the Interface

Location: L209

B2 Topic: Recent Advances in Survival Analysis

Location: L206

B3 Topic: Inequalities in Probability and Statistics

Location: L203

B4 Topic: Over-dispersion and Measurement Error

Location: L201

B5 Topic: Statistical Genetics

Location: L210

Contributed Sessions

B6 Topic: Statistical Genetics and Computational Statistics

Location: L214

B7 Topic: Time series, Applied Probability

Location: L212

Student Paper Competition Sessions

B8 Topic: Network/Clinical Trial

Location: L211

B9 Topic: Bayesian, statistical inference

Location: L213

5:30pm **Closing Session**

Location: Bakst Auditorium, BUSM Instructional Building L

Short Courses

Short Course 1 **Adaptive Clinical Trial Design and Simulation**

Instructor: Mark Chang
Executive Director, AMAG Pharmaceuticals
1:00-5:00pm, April 20, 2012, Crosstown Center Room 460/462

Summary: In this short course, we will review the basic concepts and methods for adaptive clinical trial designs, including the group sequential, sample-size re-estimation, dose-escalation, and dose-finding trials. Commonly used statistical methods for adaptive design will be introduced and compared, including the error spending approach, various methods using combinations of stagewise p-values. We will discuss implementations of adaptive trials, including interim monitoring, dynamic randomization, and analyses of adaptive trials. Practical examples using SAS and ExpDesign Studio will be provided. The FDA guidance on adaptive clinical trial designs and challenges will be discussed with recommendations. After the class, the attendees are expected to have basic knowledge to start his/her own adaptive trial design with confidence.

Biography: Dr. Mark Chang, the executive director, leads the Department of Biostatistics and Data Management with 16 years of experience as a statistician in the field of clinical trials. In addition, he has over 4 years of teaching experience as assistant professor. Before joining AMAG, Chang held various positions in Millennium Pharmaceuticals, including Director of Biostatistics and Scientific Fellow. He is a co-founder of the International Society for Biopharmaceutical Statistics, an executive member of the ASA Biopharmaceutical Section, and a member of the Expert Panel for the Networks of Centres of Excellence (NCE), Canada. He is a co-chair of the Biotechnology Industry Organization (BIO) Adaptive Design Working Group and member of the PhRMA Adaptive Design and Biomarker Working Groups. Dr. Chang is an associate editor for Statistic Journals and has over 40 publications including five books. He also serves on the Editorial Boards for the Journal of Biopharmaceutical Statistics, Statistics in Biopharmaceutical Research (ASA Journal), and the Open Public Health Journal. He has been invited to serve as a co-chair on the scientific advisory and organization committees for national and international professional/academic conferences on statistics and clinical trial designs. He has edited special issues for Journal of Biopharmaceutical Statistics, discussing the FDA guidance (Draft) on adaptive designs and has been invited twice to present statistical topics to the US Food and Drug Administration. He was invited by international medical journals to write opinion papers on clinical trials. He has taught over ten statistical short courses recently. He was recently interviewed by Journalists from the Nature Group and other Scientific Journals on innovative trial designs. Dr. Chang is an adjunct professor of Boston University and an elected fellow of the American Statistical Association.

Short Course 2 **Statistical Analysis of Network Data**

Instructor: Eric Kolaczyk
Professor, Department of Mathematics and Statistics, Boston University
1:00-5:00pm, April 20, 2012, Crosstown Center Room 460/462

Summary: Over the past decade, the study of so-called "complex networks" that is, network-based representations of complex systems - has taken the sciences by storm.

Researchers from biology to physics, from economics to mathematics, and from computer science to sociology, are more and more involved with the collection, modeling and analysis of network-indexed data. With this enthusiastic embrace of networks across the disciplines comes a multitude of statistical challenges of all sorts - many of them decidedly non-trivial. In this short course, we will cover a brief overview of the foundations common to the statistical analysis of network data across the disciplines, from a statistical perspective, in the context of topics like network summary and visualization, network sampling, network modeling and inference, and network processes. Concepts will be illustrated drawing on examples from bioinformatics, computer network traffic analysis, neuroscience, and social networks.

Biography: Eric Kolaczyk is Professor of Statistics, and Director of the Program in Statistics, in the Department of Mathematics and Statistics at Boston University, where he also is an affiliated faculty member in the Program in Bioinformatics, the Program in Neuroscience, and the Division of Systems Engineering. Before coming to Boston University, he was faculty in the Department of Statistics at the University of Chicago. In addition, he has been a visiting faculty at Harvard University, the Universite de Paris VII, and l'Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE) in Paris. Prof. Kolaczyk's main research interests currently revolve around the statistical analysis of network-indexed data, and include both the development of basic methodology and inter-disciplinary work with collaborators in bioinformatics, computer science, geography, neuroscience, and sociology. Besides various research articles on these topics, he has also authored a book in this area - Statistical Analysis of Network Data: Methods and Models (Springer, 2009). He has given various short courses on material from his book in recent years, including for the Center for Disease Control (CDC) and the Statistical and Applied Mathematical Sciences Institute (SAMSI) in the US as well as similar venues in Belgium, England, and France. Prior to his working in the area of networks, Prof. Kolaczyk spent a decade working on statistical multi-scale modeling. Prof. Kolaczyk has served as associate editor on several journals, including currently the Journal of the American Statistical Association. He has also served as co-organizer for workshops focused on networks and network data. He is an elected fellow of the American Statistical Association (ASA), an elected senior member of the Institute for Electrical and Electronics Engineers (IEEE), and an elected member of the International Statistical Institute (ISI).

Short Course 3 **Stochastic Modeling of Limit Order Books: a Journey Across Time Scales**
Instructor: Rama Cont
Professor, Columbia University
1:00-5:00pm, April 20, 2012, Crosstown Center Room 460/462

Summary: An increasing proportion of financial transactions take place in electronic markets where buy and sell orders submitted by market participants are centralized in a limit order book and executed according to precise time and price priority rules. The availability of (TeraBytes of) high-frequency data on limit order books offer a fascinating glimpse into the dynamics prices, supply and demand in financial markets and pose interesting challenges in terms of statistical modeling, both for market participants and for those - regulators and economists - who seek to understand the consequences of high frequency trading. This course will serve as an introduction to the statistical modeling of limit order books: after describing the nature of the data and the time scales involved and reviewing some of the statistical properties of

limit order books, we will argue that a limit order book has a natural description in terms of a spatial point process or queueing system, and provide various example of point process models proposed in the recent literature. Applications of such models involve time scales ranging from the millisecond (interval between orders) and the day (time needed to liquidate a large batch of shares). In the second part of the lectures, we show how functional limit theorems may be used as a useful tool to link high-frequency behavior of order flow to features such as price volatility and autocorrelation of price movements at lower frequencies. We will use fluid limits and functional central limit theorems to shows that, in liquid markets where orders arrive with high frequency, the dynamics of buy and sell queues may be approximated by a Markovian jump-diffusion process. This approximation provides an analytically tractable description of the dynamics of the order book and the market price and yields a quantitative link between statistical properties of the price process and properties of the order flow. Finally, we will sketch some open problems and challenges posed by large high-frequency data sets and discuss the potential for statistical learning methods for studying these issues.

Biography: Rama Cont is Associate professor at Columbia University (New York), director of the Columbia Center for Financial Engineering and CNRS Research Scientist at Laboratoire de probabilites (Universite de Paris VI). His research deals with stochastic analysis and stochastic modeling of financial risks, with a focus on the modelling of extreme risks - market discontinuities, systemic risk and endogenous risk - in financial markets. He was awarded the Louis Bachelier Prize by the French Academy of Sciences in 2010 for his research on mathematical modeling in finance. He is co-author of Financial Modeling with Jump Processes (CRC Press, 2003) and the Editor-in-chief of the Encyclopedia of Quantitative Finance (Wiley, 2010) and has served as a consultant to numerous financial institutions and regulatory bodies in Europe and the US.

Featured Keynote Speakers

Session 1 **Branching Process Models of Cancer**

Speaker: Rick Durrett
Duke University
10:00-11:00am, April 21, 2012, Bakst Auditorium, Instructional Building L

Abstract: It is common to use a multitype branching process to model the accumulation of mutations that leads to cancer progression, metastasis, and resistance to treatment. In this talk I will describe results from multitype branching processes that are useful in evaluating possible screening strategies for ovarian cancer, and in quantifying the amount of heterogeneity in a tumor.

Biography: Richard (Rick) Durrett received his PhD in Operations Research from Stanford University in 1976. After tens years at UCLA, he joined the Department of Mathematics at Cornell University, where he remained until recently, when he moved to Duke University. Durrett is a probabilist who has made fundamental contributions to a variety of areas, with a particular emphasis on applications in biology. There his work has involved the development of models related to spatial ecology, genetics, and, most recently, cancer biology. He is the author of nine books and close to 200 papers, and has produced over 40 PhD students. He is a past editor of the Annals of Applied Probability. Among the many honors he has received are Sloan, AMS Centennial, and Guggenheim Fellowships, election as fellow of the Institute of Mathematical Statistics (IMS) and the American Academy of Arts and Sciences (AAAS), and election to the National Academy of Science.

Session 2 **The Central Role of Modern Regression in Statistical Thinking about Neural Spike Trains**

Speaker: Robert Kass
Carnegie Mellon University
2:30-3:30pm, April 21, 2012, Bakst Auditorium, Instructional Building L

Abstract: One of the most important techniques in learning about the functioning of the brain has involved examining neural activity in laboratory animals under differing experimental conditions. Neural information is represented and communicated through series of action potentials, or spike trains, which are represented probabilistically as point processes. Because repeated presentations of stimuli often produce quite variable neural responses, statistical models have played an important role in advancing neuroscientific knowledge. In my talk I will outline some of the progress made, by many people, over roughly the past 10 years using point process regression models, and I will highlight recent work on neural synchrony (Kass, Kelly, and Loh, 2011, Annals of Applied Statistics). I will also use this body of work as a starting point for remarks about the central role of regression in statistical thinking more generally.

Biography: Robert E. (Rob) Kass received his Ph.D. in Statistics from the University of Chicago in 1980. His early work formed the basis for his book Geometrical Foundations of Asymptotic Inference, co-authored with Paul Vos. His subsequent research has been in Bayesian inference and, most recently, in the application of statistics to neuroscience. Kass is known not only for his methodological contributions, but also for several major review articles, including one with Adrian Raftery

on Bayes factors (JASA, 1995) one with Larry Wasserman on prior distributions (JASA, 1996), and a pair with Emery Brown on statistics in neuroscience (Nature Neuroscience, 2004, also with Partha Mitra; J. Neurophysiology, 2005, also with Valerie Ventura). Brown and Kass have recently attempted to stir debate about statistical education in an article entitled "What is Statistics?" (American Statistician, 2009). Kass has served as Chair of the Section for Bayesian Statistical Science of the American Statistical Association, Chair of the Statistics Section of the American Association for the Advancement of Science, Executive Editor of the international review journal Statistical Science, and founding Editor-in-Chief of the journal Bayesian Analysis. He is an elected Fellow of the American Statistical Association, the Institute of Mathematical Statistics, and the American Association for the Advancement of Science. He has been recognized by the Institute for Scientific Information as one of the 10 most highly cited researchers, 1995-2005., in the category of mathematics. In 1991 he began the series of workshops Case Studies in Bayesian Statistics, which are held at Carnegie Mellon every odd year, and was co-editor of the six proceedings volumes that were published by Springer. He is coorganizer of the workshop series Statistical Analysis of Neuronal Data, which began in 2002 and is held at Carnegie Mellon on even years. Kass has been been on the faculty of the Department of Statistics at Carnegie Mellon since 1981 and served as Department Head from 1995 to 2004; he joined the Center for the Neural Basis of Cognition in 1997, and the Machine Learning Department in 2007.

Detailed Program of Parallel Sessions

Session A1 **Recent Advances in Multivariate and Functional Data Analysis**
(Invited Session) L203 11:30-1:00pm
Organizer: Surajit Ray

Talks:

1. Finite Mixture Models for Biomedicine at the Single Cell Level
Saumyadipta Pyne
Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School

2. Identification and Estimation in Semiparametric Mixtures
Daniel Hohmann, Hajo Holzmann
Marburg University, Germany

3. Functional Factor Analysis for Periodic Remote Sensing Data
Chong Liu(1) Surajit Ray(1) Giles Hooker(2) Mark Friedl(1)
(1)Boston University (2)Cornell University

Session A2 **Panel Discussion**
Topic: Teaching Business Statistics: Best Wisdom and Biggest Challenges
Organizer: Erol Pekoz, L201 11:30-1:00pm

Discussants: John McKenzie, Tevfik Aktekin, Keith Ord and Erol Pekoz

Session A3 **Statistics and Clinical Trials**
(Invited Session) L210 11:30-1:00pm
Organizer: Gheorghe Doros

Talks:

1. Interim Sample Size Reassessment to Fixed Duration Trials
Joe Massaro (1) and Alison Pedley (2)
(1) Department of Mathematics and Statistics, Boston University, (2) Merck

2. Considerations for a Career as a Statistician in the Pharmaceutical Industry
Ronald Menton
Pfizer

3. Network Mega Analysis: Application to Historically Controlled Clinical Trials
A James O'Malley
Department of Health Care Policy, Harvard Medical School

Session A4 **Sampling and Inference from Ignorable and Non-Ignorable Network Sampling Designs**
(Invited Session) L206 11:30-1:00pm
Organizer: Edo Airoldi

Talks:

1. New Methods for Inference from Respondent-Driven Sampling Data
Krista J. Gile
Department of Mathematics and Statistics, University of Massachusetts, Amherst

2. Model-Based Estimation for Respondent-Driven Sampling
Sergiy Nesterko
Harvard University

3. Sampling on Networks: A Model-Based Approach
Simon Lunagomez
Harvard University

Session A5 **Machine Learning**
(Invited Session) L209 11:30-1:00pm
Organizer: Cynthia Rudin

Talks:

1. Ordered Rules for Classification
Dimitris Bertsimas, Allison Chang, Cynthia Rudin
Massachusetts Institute of Technology

2. Simultaneous Dimension Reduction and Variable Selection in Multivariate Regression
Lisha Chen(1), Jianhua Huang(2)
(1) Yale University (2) Texas A&M University

3. Spectral Methods for Learning Graphical Models
Sham Kakade
Microsoft Research, New England

Session A6 **Network Analysis and Covariance Matrix Estimate**
(Contributed Paper Session) L214 11:30-1:00pm
Chair: Luis Carvalho

Talks:

1. Multi-Factor Social Model for Cuda-based Bayes Estimator
Alan Lenarcic(1), Edoardo Airoldi(2), William Valdar(1)
(1)UNC Genetics, (2)Harvard Statistics

2. Estimating Network Degree Distributions from Sampled Networks: An Inverse Problem
Yaonan Zhang(1), Eric D. Kolaczyk(1) and Bruce D. Spencer(2)
(1) Department of Mathematics and Statistics, Boston University (2) Department of Statistics, Northwestern University.

3. Bayesian Degree-corrected Stochastic Block Models for Community Detection
Lijun Peng, Luis E. Carvalho
Department of Mathematics and Statistics, Boston University

4. Risky Business: Considering the Risk Set in the Business of Generative Dynamic Network Model Estimation
Alexander D'Amour, Edoardo Airoldi
Department of Statistics, Harvard University

5. Comparing Block Kronecker and Unstructured Covariance Matrix Estimation in a Hierarchical Model for Health Care Quality
Laura A Hatfield, Alan M Zaslavsky
Department of Health Care Policy, Harvard Medical School

Session A7 **Applied Statistics**
(Contributed Paper Session) L212 11:30-1:00pm
Chair: Robert Smith

Talks:

1. A Class of Discrete Transformation Survival Models with Application to Default Probability Prediction
A. Adam Ding(1) Shaonan Tian(2), Yan Yu(2) and Hui Guo(2)
(1)Northeastern University, (2)University of Cincinnati

2. Nutritional Label Use and Causal Effects on Body Mass Index
Michael Lopez, MS Roee Gutman
Department of Biostatistics, Brown University

3. Modeling Mortality and Longevity Risk
Mary M. Louie Greta M. Ljung
AIR Worldwide Corporation

4. Introducing the Hierarchical Spherical Radial R Package
Jacob Gagnon (1), Anna Liu(2)
(1) Worcester Polytechnic Institute (2) UMASS Amherst

5. Aspects of Presidential Voting (and Implications for 2012)
Robert B. Smith
Social Structural Research Inc.

Session A8 **Statistical Genetics and Bioinformatics**
(Student Paper Competition Session) L211 11:30-1:00pm
Chair: Josée Dupuis

Talks:

1. Template-Based Methods for Analyzing Chromatin Structure Dynamics Genome-Wide
Alexander W Blocker (1), Edoardo M Airoldi (1,2)
(1) Harvard University (2) The Broad Institute of MIT & Harvard

2. Identifying the Most Informative Base Pairs to Characterize the Posterior Space of RNA Secondary Structure
Luan Lin, Charles Lawrence
Division of Applied Mathematics, Brown University

3. Probabilistic Refinement of Cellular Pathway Models
Alexander Franks, Edo Airoldi
Department of Statistics, Harvard University

4. A Novel Method to Include Phenotyped but Ungenotyped Relatives in Genetic Association Tests
Wei Vivian Zhuang(1), Ching-Ti Liu(1), Gyungah Jun(1), Kathryn L. Lunetta(1,2,3)
(1)Department of Biostatistics, Boston University School of Public Health (2)Sections of General Internal Medicine, Preventive Medicine and Epidemiology, Department of Medicine, Boston University School of Medicine (3)National Heart, Lung and Blood Institute Framingham Heart Study

5. Absolute Protein Quantitation: Inference with Non-Ignorable Missing Data in High Throughput Proteomics
Eric Solis(1,2,3,4) and Edoardo Airoldi(2,3,4)
(1)Harvard University Program in Systems Biology (2)Harvard University Department of Statistics (3)Harvard FAS Center for Systems Biology (4)The Broad Institute of MIT and Harvard

Session A9 **Applied Probability**
(Student Paper Competition Session) L213 11:30-1:00pm
Chair: Henry Lam

Talks:

1. A Point Process Model for The High-Frequency Dynamics of a Limit Order Book
Allan Andersen, Rama Cont, Ekaterina Vinkovskaya
Department of Finance, Copenhagen Business School IEOR Department, Columbia University Department of Statistics, Columbia University

2. Optimal Stopping Time for the Last Passage Time and Maximum Time
Dan Ren, Kostas Kardaras
Department of Mathematics and Statistics, Boston University

3. Do Jumps Lead to Clusters?
Karthik Bharath, Vladimir Pozdnyakov and Dipak. K. Dey
Department of Statistics, University of Connecticut

4. Capturing Semantic Content with Word Frequency and Exclusivity
Jonathan Bischof and Edoardo Airoldi
Department of Statistics, Harvard University

Poster Session **Undergraduate Student Poster Competition Session** L-hallway, 2nd Floor 1:00-2:30pm

1. Assessing Value-Added Models (VAM) for teacher effectiveness
Cortley Logan and Kathleen Zarnitz
Stonehill College

2. Modeling Neural Spiking Activity in Patients with Parkinson's Disease During Movement and at Rest
Thien Tai T. Nguyen Uri T. Eden
Boston University, Mathematics and Statistics Department Undergraduate Research Opportunity Program (UROP)

Session B1 **Operations Research and Statistics: At the Interface**
(Invited Session) L209 4:00-5:30pm
Organizer: Henry Lam / Erol Pekoz

Talks:

1. Generalized Kiefer Process and Queues
Guodong Pang(1) Ward Whitt(2)
(1)Penn State University (2)Columbia University

2. Robust Risk Measurement and Model Risk
Paul Glasserman and Xingbo Xu
Columbia University

3. Dynamic Portfolio Execution
Chun Wang
Department of IE and OR, Columbia University

Session B2 **Recent Advances in Survival Analysis**
(Invited Session) L206 4:00-5:30pm
Organizer: Sangwook Kang / Jun Yan

Talks:

1. Joint Modeling of Survival Time and Longitudinal Outcomes with Flexible Random Effects
Jaeun Choi(1), Jianwen Cai(2), Donglin Zeng(2), Andrew F. Olshan(2)
(1) Harvard Medical School; (2) University of North Carolina at Chapel Hill

2. Omnibus Risk Assessment via Accelerated Failure Time Kernel Machine Modeling
Jennifer A. Sinnott and Tianxi Cai
Department of Biostatistics, Harvard School of Public Health

3. Fast Accelerated Failure Time Modeling for Case-Cohort Data
Sy Han Chiou(1), Sangwook Kang(1), and Jun Yan(1,2,3)
(1) Department of Statistics, University of Connecticut (2) Institute for Public Health Research, University of Connecticut Health Center (3) Center for Environmental Sciences & Engineering, University of Connecticut

4. Model Selection for Cox Models with Time-varying Coefficients
Jun Yan and Jian Huang
University of Connecticut and University of Iowa

Session B3 **Inequalities in Probability and Statistics**
(Invited Session) L203 4:00-5:30pm
Organizer: Mokshay Madiman

Talks:

1. Optimalities in Estimation of Large Precision Matrices
Harrison Zhou
Department of Statistics, Yale University

2. On Combinatorial Width and Uniform Donsker Classes
Ramon van Handel
Princeton University

3. Insufficiency and the Preservation of Fisher Information
David Pollard
Yale University

Session B4 **Over-dispersion and Measurement Error**
(Invited Session) L201 4:00-5:30pm
Organizer: Krishna Saha

Talks:

1. Testing for the Homogeneity of the Dispersions in the Clustered Count Data
Krishna K. Saha
Department of Mathematical Sciences, Central Connecticut State University

2. Profile Likelihood Based Confidence Interval for the Difference Between Two Correlated Proportions
Vivek Pradhan, Krishna K Saha, Tathagata Banerjee, Yuanye Zhang,and John C Evans
(1) Boston Scientific Corporation 2. Department of Mathematical Sciences, Central Connecticut State University 3. Indian Institute of Management, India 4. Department of Statistics, University of Connecticut

3. Confidence Intervals for the Dispersion Parameter in the Clustered Count Data
Debaraj Sen(1), Krishna K. Saha(2), Chun Jin(2)
(1) Department of Mathematics and Statistics, Concordia University, Canada (2) Department of Mathematical Sciences, Central Connecticut State University

4. A Functional Method for Conditional Logistic Regression with Errors-in-Covariates
Samiran Sinha
Texas A&M University

Session B5 **Statistical Genetics**
(Invited Session) L210 4:00-5:30pm
Organizer: Josée Dupuis / Ching-Ti Liu

Talks:

1. Bayesian Analysis of Gene-Environment Interactions with Error in Measurement of Environmental Covariates and Missing Genetic Data
Iryna Lobach(1), Bani Mallick(2), Raymond Carroll(2)
(1)New York University, (2)Texas A&M University

2. Mathematical Modeling for Dendritic Cells in the Immune System
Xiangtao Liu(1), Kang Liu(2), Alan Wu(1), Michel Nusenzweig(3), Anita Wang(1), Josephine Hoh(1)
(1)Yale University; (2) Columbia Univeristy; (3) Rockefeller University

3. Mixed Modeling of Meta-Analysis P-values (MixMAP) Identifies Multiple Gene Loci for Low Density Lipoprotein Cholesterol
A. S. Foulkes(1), U. Das(1), J.F. Ferguson(2), G. Matthews(1), R. Lin(1) and M.P. Reilly(2)
(1) Division of Biostatistics, University of Massachusetts, Amherst (2) Cardiovascular Institute, Perelman School of Medicine at the University of Pennsylvania

Session B6 **Statistical Genetics and Computational Statistics**
(Contributed Paper Session) L212 4:00-5:30pm
Chair: Luis Carvalho

Talks:

1. New Statistical Approaches to Missing Heritability
Or Zuk(1), Eliana Hechter(1), Shamil Sunyaev(1,2), Eric Lander(1)
(1) Broad Institute of MIT and Harvard (2) Brigham and Women's Hospital, Harvard Medical School

2. A Gene-SNP Hierarchical Bayesian Model for Genome-Wide Association Studies
Ian Johnston, Luis E. Carvalho
Department of Mathematics and Statistics, Boston University

3. Bayesian Centroid Estimation for De-Novo Motif Discovery
Luis E. Carvalho
Department of Mathematics and Statistics, Boston University

4. MODIS Land Cover Classification using Mutual Information Spanning Trees
Hunter Glanz, Luis Carvalho
Department of Mathematics and Statistics, Boston University

Session B7 **Time Series, Applied Probability**
(Contributed Paper Session) L214 4:00-5:30pm
Chair: Mamikon Ginovyan

Talks:

1. Wavelet-based Testing For Serial Correlation of Unknown Form Using Fan's Adaptive Neyman Method
Linyuan Li(1), Shan Yao(1) and Pierre Duchesne(2)
(1) University of New Hampshire, (2) University of Montreal

2. Incidental Parameter Problems Arising from the Analysis of Nonstationary Neural Point Processes
Matthew T Harrison
Division of Applied Mathematics, Brown University

3. Asymptotic Behavior of the Prediction Error for Stationary Models with Memory
Mamikon Ginovyan
Boston University

4. Order Statistics Probability Rates and Some New Results for Statistical Inference for Queues
Lee Jones
Department of Mathematical Sciences, University of Massachusetts at Lowell

5. On Fan's Adaptive Neyman Tests for Comparing Two Spectral Densities
Kewei Lu Linyuan Li
Department of Mathematics and Statistics, University of New Hampshire

Session B8 **Networks/Clinical Trials**
(Student Paper Competition Session) L211 4:00-5:30pm
Chair: Eric Kolaczyk

Talks:

1. Graphlet Decomposition of a Weighted Network
Hossein Azari Soufiani, Edoardo M. Airoldi
Department of Statistics, Harvard University

2. Uncertainty Propagation from Network Inference to Characterization
Weston Viles, Prakash Balachandran, Eric D. Kolaczyk
Department of Mathematics and Statistics, Boston University

3. Multivariate B-value: a Tool for Monitoring Data with Multiple Co-Primary Endpoints
Yansong Cheng(1), Surajit Ray(1) and Ying Zhu(2)
(1) Boston University (2) Biogen Idec

4. Identifiability of Subgroup Causal Effects in Randomized Experiments with Non-ignorable Missing Covariates
Peng Ding(1) and Zhi Geng(2)
(1) Department of Statistics, Harvard University, (2) China and Center for Statistical Science, Peking University

Session B9 **Bayesian, Statistical Inference**
(Student Paper Competition Session) L213 4:00-5:30pm
Chair: Michael LaValley

Talks:

1. Variable Selection for High-Dimensional Multivariate Outcomes with Application to Genetic Pathway/Network Analysis
Tamar Sofer(1), Lee Dicker(2) and Xihong Lin(1)
(1) Department of Biostatistics, Harvard School of Public Health (2) Department of Statistics, Rutgers University.

2. Generalized Species Sampling Priors with Latent Beta Reinforcements
Thiago Costa(1), Michele Guindani(2), Federico Bassetti(3), Fabrizio Leisen(4), Edoardo M. Airoldi(5)
(1)Department of Statistics and School of Engineering and Applied Sciences, Harvard University, (2) T. MD Anderson Cancer Center, Department of Biostatistics, (3) University of Pavia, Department of Mathematics, Italy. (4) Departamento de Estadistica, Universidad Carlos III de Madrid, Calle Madrid, Spain. (5) Department of Statistics, Harvard University

3. Bayesian Framework for the Incorporation of Multiple Data Sources with the Analysis of Infectious Disease Outbreaks
Carlee B. Moser(1), Mayetri Gupta(2), Brett N. Archer(2), and Laura F. White(1)
(1) Department of Biostatistics, Boston University School of Public Health, (2) National Institute for Communicable Diseases (NICD), a division of the National Health Laboratory Service (NHLS), South Africa

4. Analysis of Catch Counts with Excess Zeros and Over-Dispersion in Bayesian Approach
Rengui Qiao, Liliana Gonzalez
Department of Computer Science and Statistics, University of Rhode Island

5. Semiparametric Multivariate Accelerated Failure Time Model with Generalized Estimating Equations

Sy Han Chiou
Department of Statistics, University of Connecticut

Abstracts

A point process model for the high-frequency dynamics of a limit order book

Allan Andersen, Rama Cont, and Ekaterina Vinkovskaya
Department of Finance, Copenhagen Business School IEOR Department, Columbia University Department of Statistics, Columbia University

The statistical properties of events affecting a limit order book—market orders, limit orders and cancellations—reveal strong evidence of clustering in time, significant cross-correlation across event types and significant dependence of the order flow on the bid-ask spread. We show that these dependencies may be adequately represented by a multi-dimensional self-exciting point process, for which a tractable parameterization is proposed. Using high-frequency data from the Trades and Quotes database, we perform a Maximum Likelihood Estimation of the model and assess its predictive performance for a variety of stocks.
Keywords: Hawkes, self-exciting, limit order book, point process

Ordered Rules for Classification

Dimitris Bertsimas, Allison Chang, and Cynthia Rudin
Massachusetts Institute of Technology

We aim to design classifiers that are interpretable to human experts, yet match the predictive power of top machine learning algorithms (like SVMs and boosted decision trees). We propose a novel mixed integer optimization (MIO) approach for this task. Our algorithm builds decision list classifiers that are accurate, simple and insightful.
Keywords: association rules, decision trees, decision lists, interpretable models, classification

Do Jumps Lead to Clusters?

Karthik Bharath, Vladimir Pozdnyakov, and Dipak. K. Dey
Department of Statistics, University of Connecticut

We investigate the appropriateness of regarding jumps occurring in the discretely observed semimartingales as a phenomenon which leads to clustering of the data from increments of the process. To that end we develop a clustering framework, motivated by the problem of testing for jumps in such semimartingale models, and derive its asymptotic properties under a distribution-free setup. We then propose an intuitive and easily verifiable clustering criterion, based on the Empirical Cross-over Function, which provides us with the requisite tools to develop a test for the presence of jumps. We illustrate the validity of our theory on the popular Merton and Kou models for options pricing with the objective of presenting our approach as a first step towards viewing jumps in diffusion models as a clustering mechanism
Keywords: Clustering, Limit theorems, Semimartingales

Capturing semantic content with word frequency and exclusivity

Jonathan Bischof and Edoardo Airoldi
Department of Statistics, Harvard University

An ongoing challenge in the analysis of document collections is how to summarize content in terms of a set of themes or topics. The current practice of constructing topics in terms of their most frequent words limits interpretability by ignoring the differential use of words across topics. Rather, we find that words that are both common and relatively exclusive to a topic are more effective at communicating semantic content concisely, than words common across many topics. To substantiate this claim, we consider data settings where professional editors have annotated documents to a pre-specified collection of topics. Topics are organized into a tree, with the leaf-nodes corresponding to the most specific topics. Each document may be annotated to multiple topics, at different levels on the tree. Here, we introduce a Hierarchical Poisson Convolution (HPC) model to analyze annotated documents in these settings. Using HPC we infer the differential use of words across topics as well as word frequency within topics. HPC leverages known hierarchical structure on human-labeled topics to make focused comparisons of

differential usage within each branch of the tree, inferring a clear semantic description of human-generated topics in terms of words that are both frequent and exclusive. We develop a parallelized Hamiltonian Monte Carlo sampler that the inference to scale to millions of documents. We anticipate that learning a concise semantic description for any collection of topics implicitly defined by professional editors is the first step toward the semi-automated creation of domain-specific topic ontologies. Domain-specific topic ontologies may be useful for evaluating the semantic content of inferred topics, or for predicting the semantic content of new social media, including Twitter messages and Facebook wall-posts.
Keywords: Hierarchical Bayes; Hamiltonian Monte Carlo; Markov Chain Monte Carlo; Text analysis; Semantic content summarization; Topic models; Supervised learning.

Template-based methods for analyzing chromatin structure dynamics genome-wide
Alexander W Blocker (1) and Edoardo M Airoldi (1,2)
(1) Harvard University, (2) The Broad Institute of MIT and Harvard
We consider the problem of estimating chromatin structure dynamics from paired-end sequencing data. We introduce a novel approach based on nonparametric templates for modeling variability along the sequence of read counts associated with nucleosomal DNA due to enzymatic digestion and other sample preparation steps. We combine it with a more traditional generalized linear model that accounts for the variability of read counts at each genomic location due to incomplete sequencing. Variability in coverage along the genome is accounted for using a hierarchical model specification, in which segments with homogeneous coverage that approximately correspond to open reading frames and intergenic regions are conditionally independent. We evaluate the template-based approach by analyzing the sensitivity of estimates to a key smoothing parameter; its statistical power in detecting the locations of nucleosomes that are well-positioned within the population, and differentially positioned in two or more sub-populations; and the reproducibility of estimated locations of well-position nucleosomes across multiple data sets. Inference is carried out via an EM algorithm that leverages a sophisticated approximation strategy for the estimand of interest. We provide MPI Python implementations, stand-alone and on Amazon EC2, which fit the yeast genome in the order of minutes and can scale to mammalian-sized genomes within an hour. The template-based approach we develop in this paper is applicable to single-end sequencing data using alternative sources of fragment length information (Bioanalyzer), and to ordered and sequential data more generally. It provides a flexible and scalable alternative to mixture models, hidden Markov models, and Parzen window methods.
Keywords: expectation-maximization; Laplace approximation; deconvolution; detection; screening; massive data; yeast; nucleosomes; parallel computation; measurement error

Bayesian centroid estimation for de-novo motif discovery
Luis E. Carvalho
Department of Mathematics and Statistics, Boston University
Biological sequences may contain patterns that signal important biomolecular functions; a classical example is regulation of gene expression by transcription factors that bind to specific patterns in genomic promoter regions. In de-novo motif discovery we are given a set of sequences that share a common motif and aim to identify not only this motif composition, but also the binding sites in each sequence of the set. We present a Bayesian model that is an extended version of the model adopted by the Gibbs motif sampler, and propose a new centroid estimator that arises from a refined and meaningful loss function for binding site inference. We discuss the main advantages of centroid estimation for motif discovery, including computational convenience, and how its principled derivation offers further insights about the posterior distribution of binding site configurations. We also illustrate the proposed approach on both simulated and real datasets, and conclude with directions for future work.
Keywords: Gibbs sampling, stochastic backtracking

Simultaneous Dimension Reduction and Variable Selection in Multivariate Regression
Lisha Chen (1) and Jianhua Huang (2)

(1) Yale University, (2) Texas A and M University
In this talk, we will introduce a new method to address the problem of predicting several response variables from the same set of predictor variables using linear regression. The method incorporates the interrelation between the response variables to improve the overall predictive accuracy. When the dimension of the predictors is high, the new proposal conducts variable selection and dimension reduction simultaneously. We will discuss the asymptotic consistency of the proposed method. The new procedure is compared with several previously proposed variable selection methods for multivariate regression and exhibits improved accuracy in prediction and variable selection.

Multivariate B-value: a tool for monitoring data with multiple co-primary endpoints

Yansong Cheng (1), Surajit Ray (1), and Ying Zhu (2)
(1) Boston University,(2) Biogen Idec
Many clinical trials are required to declare significant efficacy on two or more primary endpoints simultaneously. Multiple co-primary endpoints problem is also called reverse multiplicity problem. Intersection-Union Test (IUT) is the standard test for such problem. One of the main questions of reverse multiplicity problem is how to conduct the interim analysis. This paper extends the B-value to multiple dimensions that makes the B-value tool be applicable for the nonsequentially designed clinical trial study with considering multiple co-primary endpoints. The correlation among the co-primary endpoints are taken into consideration. The overall conditional power (CP) is defined as the probability of declaring all endpoints are significant at the end of study, conditioning on the observed information at interim analysis. The overall CP of the IUT is no greater than the smallest CP of all sub-hypothesis and it falls between the two extreme cases: one is all the endpoints are independent and the other is all the endpoints are perfectly correlated. The example of comparing two samples with two normally co-primary endpoints is shown.

Fast Accelerated Failure Time Modeling for Case-Cohort Data

Sy Han Chiou (1), Sangwook Kang (1), and Jun Yan (1,2,3)
(1) Department of Statistics, University of Connecticut, (2) Institute for Public Health Research, University of Connecticut Health Center, (3) Center for Environmental Sciences and Engineering, University of Connecticut
Semiparametric accelerated failure time (AFT) models directly relate the predicted failure times to covariates and are a useful alternative to models that work on the hazard function or the survival function. For case-cohort data, much less development has been done with AFT models. In addition to the missing covariates outside of the sub-cohort and cases, challenges from AFT model inferences with full cohort are retained. The regression parameter estimator is hard to compute because the most widely used rank-based estimating equations are not smooth, Further, its variance depends on the unspecified error distribution, and most methods rely on computing intensity bootstrap to estimate it. We propose a fast rank-based inference procedure for AFT models, applying recent methodological advances to the context of case-cohort data. Parameters are estimated with an induced smoothing approach that smooths the estimating functions and facilitates the numerical solution. Variance estimators are obtained through an efficient resampling methods for nonsmooth estimating functions that avoids full blown bootstrap. Simulation studies suggest that the recommended procedure provides fast and valid inferences among several competing procedures. Application to a tumor study demonstrates the utility of the proposed method in routine data analysis.
Keywords: induced smoothing; multiplier bootstrap; resampling; stratified sampling

Semiparametric Multivariate Accelerated Failure Time Model with Generalized Estimating Equations

Sy Han Chiou
Department of Statistics, University of Connecticut
The semiparametric accelerated failure time model is not as widely used as the Cox relative risk model mainly due to computational difficulties. Recent developments in least squares estimation and induced

smoothing estimating equations provide promising tools to make the accelerate failure time models more attractive in practice. For semiparametric multivariate accelerated failure time models, we propose a generalized estimating equation approach to account for the multivariate dependence through working correlation structures. The marginal error distributions can be either identical as in sequential event settings or different as in parallel event settings. Some regression coefficients can be shared across margins as needed. The initial estimator is a rank-based estimator with Gehan's weight, but obtained from an induced smoothing approach with computation ease. The resulting estimator is consistent and asymptotically normal, with a variance estimated through a multiplier resampling method. In a simulation study, our estimator is up to three times as efficient as the initial estimator, especially with stronger multivariate dependence and heavier censoring percentage. Two real examples demonstrate the utility of the proposed method.

Joint Modeling of Survival Time and Longitudinal Outcomes with Flexible Random Effects

Jaeun Choi (1), Jianwen Cai, Donglin Zeng, and Andrew F. Olshan (2)
(1) Harvard Medical School, (2) University of North Carolina at Chapel Hill

In biomedical or public health research, it is common for both survival time and longitudinal outcomes to be collected for a subject, along with the subject's characteristics or risk factors. Joint analysis of longitudinal outcomes and survival time is used to find important variables for predicting both longitudinal outcomes and survival time which are correlated within a subject. Random effects are introduced to account for dependence between survival time and longitudinal outcomes due to unobserved factors. Gaussian distribution is conventionally assumed for random effects. Misspecifying normality assumption can lead to serious bias in estimation. We relax the normality assumption of random effects by assuming the underlying distribution to be unknown. We propose a mixture of Gaussian distributions as an approximation in estimation. Weights of the mixture components are estimated with model parameters using the EM algorithm. The observed information matrix is adopted to estimate the asymptotic variances of the proposed estimators. The method is demonstrated to perform well in finite samples via simulation studies. We also conduct simulation studies to examine the robustness of the mixture distribution. AIC and BIC criteria are adopted for selecting the number of mixtures, and the selection procedures are assessed through simulation studies. We illustrate our approach with the data from the Carolina Head and Neck Cancer Study.

Keywords: Gaussian mixtures; Generalized linear mixed model; Maximum likelihood estimator; Random effect; Simultaneous modeling; Stratified Cox proportional hazards model

Generalized Species Sampling Priors with Latent Beta reinforcements

Thiago Costa (1), Michele Guindani (2), Federico Bassetti (3), Fabrizio Leisen (4), and Edoardo M. Airoldi (5)
(1) Department of Statistics and School of Engineering and Applied Sciences, Harvard University, (2) T. MD Anderson Cancer Center, Department of Biostatistics, (3) University of Pavia, Department of Mathematics, (4) Departamento de Estadistica, Universidad Carlos III de Madrid, (5) Department of Statistics, Harvard University

Many popular Bayesian Nonparametric priors can be characterized in terms of exchangeable species sampling sequences. However, in some applications, exchangeability may not be appropriate. We introduce non exchangeable generalized species sampling sequences characterized by a tractable predictive probability function with weights driven by a sequence of independent Beta random variables. We compare their clustering properties with those of the Dirichlet Process and the two parameters Poisson-Dirichlet process. We propose the use of such sequences as prior distributions in a hierarchical Bayes modeling framework. We detail on Markov Chain Monte Carlo posterior sampling and discuss the resulting inference in a simulation study, comparing their performance with that of popular Dirichlet Processes mixtures and Hidden Markov Models. Finally, we discuss an application to the detection of chromosomal aberrations in breast cancer using array CGH data.

Keywords: Bayesian non-parametrics, Species Sampling Priors, Predictive Probability Functions, Random Partitions

Risky Business: Considering the Risk Set in the Business of Generative Dynamic Network Model Estimation

Alexander D'Amour and Edoardo Airoldi
Department of Statistics, Harvard University

Understanding the dynamics of interaction networks, particularly those that evolve in time, is a topic of significant interest in a number of fields. While many latent and observed covariate generating processes have been proposed for static and dynamic networks, few are well-equipped to account for the sparsity inherent to large interaction networks, despite leveraging this sparsity to make computation more efficient. In this paper, we consider a specific model that describes dynamic network data as a Cox Proportional Hazards process that relates observed node- and dyad-level covariates to the appearance of interactions in the adjacency matrix over time. This model is particularly illustrative because, for historical reasons, it includes the specification of a *risk set* in its formulation, which makes the roles of structural zeros in the data explicit. While the risk set is largely untreated in previous applications of this model to networks, we show that risk set specification plays a critical role in parameter estimation, and that simple guesses at the composition of the risk set at best underestimate variance and at worst violate the necessary conditions for consistent parameter estimation. We choose to frame the specification of the risk set as a missing data problem and propose a data augmentation strategy, similar to zero-inflation, to account for uncertainty about the unobserved risk set while estimating the parameters of the Cox Process. Our method is computationally efficient, maintains the conditions for asymptotic consistency, and accounts for the sparsity of the observed network in a principled way. The zero-inflation principle used in this model is broadly applicable to network GLM's as a whole, even if they do not explicitly include a risk set in their specification. As the available covariates are often insufficient to explain the full sparsity pattern in a network, a zero-inflation strategy is a principled and effective way to account for residual sparsity while maintaining an easily interpretable generating process.

Keywords: dynamic network; cox process; missing data; data augmentation; zero inflation; sparse network

Identifiability of Subgroup Causal Effects in Randomized Experiments with Nonignorable Missing Covariates

Peng Ding (1) and Zhi Geng (2)
(1) Department of Statistics, Harvard University, (2) Center for Statistical Science, Peking University, China

Randomized experiments are widely regarded as the gold standard for estimating causal effects. In some cases, however, missing data can complicate the analysis, especially if the missing data mechanism is nonignorable. The missing data of the pretreatment covariate makes it challenging to estimate the subgroup causal effects, which are defined by the possibly missing covariate. When the missing data mechanism is nonignorable, the parameters of interest are generally not point identifiable, and we can only get bounds for the parameters of interest, which may be too wide for practical use. In some real cases, we have prior knowledge that some restrictions may be plausible. We show the identifiability of the causal effects and joint distributions for four interpretable models. Application of our methods to a real data shows that the nonignorable missing data model fits better than the ignorable missing data model, and the results conform to the study's original expert opinions.

Keywords: Bound; Causal inference; Missing data; Nonignorable.

A Class of Discrete Transformation Survival Models with Application to Default Probability Prediction

A. Adam Ding (1), Shaonan Tian, Yan Yu and Hui Guo (2)
(1) Northeastern University, (2) University of Cincinnati;

Accurate corporate default probability prediction is very important for banking capital reservation calculation. While the corporate default can be naturally considered as a survival event, the survival analysis theory and techniques were not used for this application until last decade. In this talk, we discuss some distinct features of bankruptcy from the traditional survival analysis model, and apply a discrete transformation family of survival analysis to corporate default risk predictions. We show on the default data of the US companies from 1980-2006 that a transformation parameter different from the popular Shumway's model and the proportional hazards model is needed for default prediction. The predicted corporate default probabilities on this data set show that the distress company stocks do not receive full risk premium as speculated by the famous Fama and French's (1996) conjecture.
Keywords: Corporate Bankruptcy Prediction; Credit Risk; Logistic Regression; Proportional Hazards; Survival Analysis.

Mixed modeling of Meta-Analysis P-values (MixMAP) identifies multiple gene loci for low density lipoprotein cholesterol
A. S. Foulkes (1), U. Das (1), J.F. Ferguson (2), G. Matthews (1), R. Lin (1), and M.P. Reilly (2)
(1) Division of Biostatistics, University of Massachusetts, Amherst (2) Cardiovascular Institute, Perelman School of Medicine at the University of Pennsylvania
Informing missing heritability for complex disease will likely require leveraging information across multiple SNPs within a gene region simultaneously to characterize gene and locus level contributions to disease phenotypes. To this aim, we introduce a novel strategy, termed Mixed modeling of Meta-Analysis P-values (MixMAP), that draws on a principled statistical modeling framework and the vast array of summary data now available from genetic association studies, to test formally for locus level association. The primary inputs to this approach are: (a) single SNP level p-values for tests of association; and (b) the mapping of SNPs to genomic regions. The output of MixMAP is comprised of locus level estimates and tests of association. In application of MixMAP to summary data from the Global Lipids Gene Consortium, we have identified multiple new loci for low-density lipoprotein cholesterol (LDL-C), a causal risk factor for cardiovascular disease and we also demonstrate the potential utility of MixMAP in small data settings. Overall, MixMAP offers novel and complementary information as compared to traditional analysis approaches and is straightforward to implement with existing open-source statistical software tools.
Keywords: SNP, GWAS, LDL-cholesterol, mixed effects modeling, meta-analysis

Probabilistic refinement of cellular pathway models
Alexander Franks and Edo Airoldi
Department of Statistics, Harvard University
Building better models of cellular pathways is one of the major challenges of systems biology and functional genomics. There is a need for methods to build on established expert knowledge and reconcile it with results of high-throughput studies. Moreover, the available data sources are heterogeneous and need to be combined in a way specific for the part of the pathway in which they are most informative. Here, we present a compartment specific strategy for integrating both edge and node data for the refinement of a network hypothesis. Specifically, we use a local-move Gibbs sampler for refining pathway hypotheses from a compendium of heterogeneous data sources, including novel methodology for integrating protein attributes. We demonstrate the utility of this approach in a case study of the pheromone response MAPK pathway in the yeast S. cerevisiae.

Introducing the Hierarchical Spherical Radial R Package
Jacob Gagnon (1) and Anna Liu (2)
(1) Worcester Polytechnic Institute, (2) UMASS Amherst
We are interested in studying the effects of zidovudine, lamivudine, and ritonavir on immune reconstitution for 48 patients enrolled in an AIDS clinical trial performed by Lederman et al. We utilize a generalized semiparametric mixed model to characterize HIV dynamics, to determine the relationship

between HIV viral load with CD4 and CD8 cell counts, and to determine the success of this antiretroviral regimen. Compared to previous work, we show an improved time complexity and an accuracy comparable with Adaptive Gaussian Quadrature. We will also discuss work in progress of developing an R package of our methods: the Hierarchical Spherical Radial R Package.
Keywords: AIDS, R package, mixed models, semi-parametric, HIV, GSMM

New methods for inference from Respondent-Driven Sampling Data
Krista J. Gile
Department of Mathematics and Statistics, University of Massachusetts, Amherst
Respondent-Driven Sampling is type of link-tracing network sampling used to study hard-to-reach populations. Beginning with a convenience sample, each person sampled is given 2-3 uniquely identified coupons to distribute to other members of the target population, making them eligible for enrollment in the study. This is effective at collecting large diverse samples from many populations. Current estimation relies on sampling weights estimated by treating the sampling process as a random walk on the underlying network of social relations. These estimates are based on strong assumptions allowing the data to be treated as a probability sample. In particular, existing estimators assume a with-replacement sample with an ideal initial sample. We introduce two new estimators, the first based on a without-replacement approximation to the sampling process, and the second based on fitting a social network model (ERGM), and demonstrate their ability to correct for biases due to the finite population and initial convenience sample. Our estimators are based on a model-assisted design-based approach, using standard errors based on a parametric bootstrap. We conclude with an application to data collected among injecting drug users, including extension to observable features of the sampling process. This talk includes joint work with Mark S. Handcock
Keywords: link-tracing, respondent-driven sampling, network, network sampling

Asymptotic Behavior of the Prediction Error for Stationary Models with Memory
Mamikon Ginovyan
Boston University
Let $X(t)$, $t = 0, \pm 1, \dots$, be a second order stationary random sequence with spectral function $F(\lambda)$, $\lambda \in [-\pi, \pi]$. Denote by $\sigma_T^2(F)$ the best linear mean square one-step prediction error variance in predicting the random variable $X(0)$ by the past of $X(t)$ of length T, and let $\sigma^2(F) = \sigma_\infty^2(F)$ be the prediction error variance by the entire past. The Kolmogorov-Szegö classical theorem states that $\sigma^2(F) = \sigma^2(f)$, and the relative prediction error defined by $\delta_T(F) := \sigma_T^2(F) - \sigma^2(f)$ is nonnegative and tends to zero as $T \to \infty$, where $f(\lambda)$ is the spectral density of $X(t)$. In this talk we will present some results that describe the rate of decrease of the relative prediction error $\delta_T(F)$ to zero as $T \to \infty$, depending on the memory structure (short-, intermediate-, and long-memory) of the underlying process $X(t)$, and the smoothness properties of its spectral density function $f(\lambda)$. We also will discuss the inverse problem: for a given rate of decrease of the relative prediction error $\delta_T(F)$ to zero, describe the model process $X(t)$ compatible with that rate. Specify then dependence structure of $X(t)$ and the smoothness properties of its spectral density $f(\lambda)$.
Keywords: Prediction Error, Stationary, Memory, Spectral Density

MODIS Land Cover Classification using Mutual Information Spanning Trees
Hunter Glanz and Luis E. Carvalho
Department of Mathematics and Statistics, Boston University
The task of land cover classification comprises a major portion of the remote sensing community and presents a natural setting for the application of graphical models defined on the lattice induced by the grid of pixels that make up a remotely sensed image. For a particular image, we observe a time series with 46 observations per pixel of an enhanced vegetation index (EVI) computed from the spectral bands captured by the Moderate Resolution Imaging Spectroradiometer (MODIS) and wish to categorize each pixel according to the International Geosphere–Biosphere Programme (IGBP) classification. To this

end, we propose a Bayesian model that incorporates the spatial co-dependencies in the lattice using a Potts model prior. However, traditional inferential approaches to classification are often computationally intractable due to the size and topology of such lattices. To overcome the computational complexity we approximate the lattice with a maximum weighted spanning tree where weights are taken as the mutual information between naive posterior distributions on neighboring pixels. The motivation in using a tree based on mutual information is to preserve the strongest spatial co-dependencies as measured by the information gain of having an edge in the tree. Given this spanning tree approximation, we show how to efficiently compute centroid estimates from the exact posterior distribution of land cover classifications and discuss our results in light of a well-accepted reference dataset.
Keywords: graphical models, centroid estimation

Robust Risk Measurement and Model Risk
Paul Glasserman and Xingbo Xu
Columbia University
Financial modeling is subject to model risk due, e.g., to model mispecification and statistical estimation uncertainty, and the risk can be amplified by combining optimization. We use relative entropy and $\alpha-$divergence to quantify model risk and study its the effect on a general problem which optimizes statistical expectation. We formulate it as a robust problem, and the worst scenario is obtained. We further provide a survey of various financial models on their vulnerability to model risk, where we compare the nominal model, its worst scenario and candidate perturbations. We find uncertainty in parameters are not adequate to describe the worst scenarios, except for some simple models.
Keywords: Model risk, relative entropy, divergence, robustness

On combinatorial width and uniform Donsker classes
Ramon van Handel
Princeton University
I will discuss a trivial combinatorial inequality for Gaussian processes that gives a surprisingly sharp necessary condition for the uniform Donsker property.

Incidental parameter problems arising from the analysis of nonstationary neural point processes
Matthew T Harrison
Division of Applied Mathematics, Brown University
The spiking dynamics of simultaneously recorded neurons from a small region of cortex reflect the local network structure of excitatory and inhibitory connections between observed neurons, as well as the time varying response of the neurons to their many unobserved and correlated inputs. Inference about the local network using overly simplified models is easily contaminated by these unobserved nonstationary influences for overly simplified models. More appropriate models suffer from challenging incidental parameter problems. Problems are illustrated and a conditional inference solution is proposed.
Keywords: point processes, binary time series, network, nonstationary, conditional inference

Comparing block Kronecker and unstructured covariance matrix estimation in a hierarchical model for health care quality
Laura A Hatfield and Alan M Zaslavsky
Department of Health Care Policy, Harvard Medical School
In hierarchical models, it may be useful to enforce some restrictions on the covariance matrix, such as autoregressive, compound symmetric, or block diagonal, according to the problem at hand. Here, we develop an approach to estimating block structured covariance matrices when some of the blocks have additional Kronecker product constraints. The motivating application concerns ratings of and reports on health care quality by Medicare beneficiaries enrolled in Medicare Advantage managed care plans. The Consumer Assessments of Health Plans Study (CAHPS) survey includes items concerning experiences

with the plan, individual physicians, prescriptions, and access to care. Previous work has found that age, general health status, and educational attainment are important predictors of quality ratings. Plans that attract enrollees that tend to provide lower quality ratings should not be penalized for this; instead, differences among plans should be standardized to a common casemix. The usual approach to casemix adjustment assumes that the effects casemix variables are the same across plans; however, this may be violated if some plans provide more differential quality of care to individuals with certain characteristics, e.g., if in some plans the less healthy beneficiaries experience worse care than healthier ones, to a greater degree than in other plans. We wish to assess variation across plans in the effects of casemix variables. To properly account for the complex missingness resulting from skip patterns in the CAHPS survey, we first fit plan-level models relating individual quality scores to age, education and health status. The resulting coefficient estimates (and their variances) constitute the first level of a hierarchical model. Our primary interest is in the second-level covariance matrix of this model, which characterizes how the coefficients vary across plans. Assuming proportional relationships among the casemix coefficients across outcomes, we can dramatically reduce the number of free parameters in the covariance matrix by imposing a block Kronecker structure. We construct an EM algorithm to estimate the covariance matrix in both the unstructured and block Kronecker structured cases, and compare the resulting estimates.

Identification and estimation in semiparametric mixtures

Daniel Hohmann and Hajo Holzmann
Marburg University, Germany

For finite mixtures where at least for some component distributions no parametric form is assumed, identification and estimation are often based on shape-constraints like symmetry or log-concavity. We consider two-component mixtures where the weights depend on additional covariates. Such models include multivariate mixtures with independent components, Markov-dependent mixtures or regression models with a misspecified binary regressor. Identification results can be obtained in case of certain tail conditions on the distribution functions or the characteristic functions of the mixture components. We propose estimators of the mixture components and, using strong approximations of the ordinary empirical and the empirical characteristic processes, we prove their asymptotic normality.

Keywords: finite mixtures, semiparametric mixtures, identifiability

A New Class of Flexible Link Function with Application to Species Co-occurrence Study

Xun Jiang and Dipak K. Dey
Deparment of Statistics, University of Connecticut

In this paper, we propose a new family of flexible link functions for modeling binomial response data. By introducing a power parameter into the cdf corresponding to a symmetric link function and its mirror reflection, greater flexibility in skewness can be achieved in both positive as well as negative directions. Through simulated data sets, we show the proposed link function performs better fitting against link misspecification than existing standard link functions by allowing data to determine the direction and magnitude of skewness. Finally, we present the Protea species co-occurrence data from Cape Floristic Region of South Africa as an illustrative example to demonstrate the effectiveness of the proposed link functions.

Keywords: Bayesian method, community ecology, generalized linear model, MCMC, model selection, power link function

A gene-SNP hierarchical Bayesian model for genome-wide association studies

Ian Johnston and Luis E. Carvalho
Department of Mathematics and Statistics, 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. To obtain posterior samples we use a Gibbs sampler which we can later extend to take advantage of particular features of the resulting graphical model to obtain an efficient sampling procedure. We propose conducting 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 discuss preliminary results on real-world data.
Keywords: Gibbs sampling, large p small n, graphical model, centroid estimation

Order statistics probability rates and some new results for statistical inference for queues
Lee Jones
Department of Mathematical Sciences, University of Massachusetts at Lowell
Efficient algorithms were initially developed for computing the probability that the order statistics of n i.i.d. uniform random variables lie in a given n-dimensional rectangular region in order to calculate the cumulative distribution of the Kolmogorov statistic. These algorithms were rediscovered and used to find expected queue length(and other queue performance measures) in a queuing system from the set of start/stop service data in a time interval in the interior of which each server who became free was immediately reengaged by a waiting customer.With practical data there are time gaps in restarting service with a waiting customer. By generalizing the order statistics probability computational problem and developing feasible algorithms for its solution we can give confidence intervals for queue performance measures for such data.
Keywords: order statistics, queue inference, Kolmogorov statistic

Spectral Methods for Learning Graphical Models
Sham Kakade
Microsoft Research, New England
This work presents a methodology for learning graphical models with hidden nodes through algebraic techniques (in particular, matrix decomposition and spectral methods), using independent samples of the observed variables. The talk focuses on tree models, and covers two aspects of the underlying learning problem: parameter estimation and structural learning. The underlying idea is to utilize the spectral decomposition of the second moment matrix to reveal the latent structure. The first part is concerned with parameter estimation. Here, we present an an efficient and provably correct algorithm for learning HMMs (i.e. recovering the correct HMM dynamics), with a sample complexity depending on some mild conditions of the underlying system. The algorithm is also simple, employing only a singular value decomposition and matrix multiplications, and does not suffer from local minimum issues in non-convex optimization (such as for more traditional approaches, including the EM algorithm), and it handles high dimensional observations and long range dependencies more easily. The method can be extended to estimating parameters for nonlinear systems and general tree structured graphical models with unobserved nodes. The second part is concerned with structural learning, where we provide the Spectral Recursive Grouping algorithm, an efficient and simple procedure for recovering the underlying tree topology of a broad class of multivariate tree models with hidden nodes. Exact recovery of the tree structure can be established based on certain natural dependencies on statistical and structural properties of the underlying joint distribution. Join work with: Daniel Hsu, Tong Zhang; Anima Anandkumar, Kamalika Chaudhuri, Le Song
Keywords: graphical models, spectral methods, hidden markov models, trees

Multi-Factor Social model for CUDA-based Bayes Estimator
Alan Lenarcic (1), Edoardo Airoldi (2), and William Valdar (1)
(1) UNC Genetics, (2) Harvard Statistics
Clustering of social networks can be difficult when they exhibit "Expander" geometries that encourage small-world paths. We explore an explanation based upon the Ahn et al. 2010 crique, and suggest

that most social actors, acting as individuals, affiliate to multiple social groups. A simplified latent probability model suggests an EM algorithm suitable for massive parallelization on graphical-processing-units (GPUs). Using the UNC BASS nVidia-CUDA cluster, we discuss the differences and pitfalls of proposing network statistics using Bayes inference developed for streamlined computation. Theoretical heavy-tailed properties of the network edge distribution are compared to networks from social and biological 2hybrid datasets.
Keywords: Networks, GPU, Latent Probability Models

Wavelet-based Testing For Serial Correlation of Unknown Form Using Fan's Adaptive Neyman Method
Linyuan Li (1), Shan Yao (1), and Pierre Duchesne (2)
(1) University of New Hampshire, (2) University of Montreal
A wavelet-based consistent test for serial correlation of unknown form is developed. The proposed test is constructed by empirical wavelet coefficients of wavelet-based spectral density estimator. The asymptotic normal distributions of the empirical wavelet coefficients under the null hypothesis of no serial correlation are derived. Furthermore, they are asymptotically uncorrelated and have a limit multivariate normal distribution. Adopting Fan's (1996) canonical multivariate normal hypothesis testing model, a new adaptive test statistics is proposed. The proposed test is completely data-driven or adaptive, which avoids the need to select any smoothing parameters. Furthermore, under a suitable class of local alternatives, our test is consistent against serial correlation of unknown form. The proposed test has better power than current tests, when the true spectral density has significant spatial inhomogeneity, such as seasonal or business cycle periodicities in economic and financial time series. In general, the convergence of the proposed test statistic toward its respective asymptotic distribution appears to be relatively slow. In view of that, Monte Carlo method is investigated to determine the corresponding critical values. In a small simulation study, several current test statistics are compared, with respect to their levels and powers.
Keywords: adaptive Neyman test; periodogram; spectral density; time series; wavelet method.

Identifying the most informative base pairs to characterize the posterior space of RNA secondary structure
Luan Lin and Charles Lawrence
Division of Applied Mathematics, Brown University
The posterior space of RNA secondary structure often shows multi modal. In this paper we construct a characterization criteria based on entropy measure and propose an iterative algorithm which identifies the crucial base pairs during the folding process as an avenue for improved characterization of the posterior space. In addition to exact calculation, a sampling based algorithm is also presented.
Keywords: clustering, mutual information, RNA secondary structure

Functional Factor Analysis for Periodic Remote Sensing Data
Chong Liu, Surajit Ray (1), Giles Hooker (2), and Mark Friedl (1)
(1) Boston University, (2) Cornell University
We present a new approach to factor rotation for functional data. This is achieved by rotating the functional principal components towards a pre-defined space of periodic functions designed to decompose the total variation into components that are nearly-periodic and nearly-aperiodic with a pre-defined period. We show that the factor rotation can be obtained by calculation of canonical correlations between appropriate spaces which makes the methodology computationally efficient. Moreover we demonstrate that our proposed rotations provide stable and interpretable results in the presence of highly complex covariance. This work is motivated by the goal of finding interpretable sources of variability in gridded time series of vegetation index measurements obtained from remote sensing, and we demonstrate our methodology through an application of factor rotation of this data.

Keywords: Factor rotation, variance decomposition, functional data analysis, covariance surface, remote sensing, principal periodic components.

Mathematical Modeling for Dendritic Cells in the Immune System

Xiangtao Liu (1), Kang Liu (2), Alan Wu (1), Michel Nusenzweig (3), Anita Wang (1), and Josephine Hoh (1)
(1) Yale University, (2) Columbia University, (3) Rockefeller University

Dendritic cells (DCs) are specialized antigen presenting cells (APCs) playing key roles in initializing immune responses. Their functions have been thoroughly studied, and many applications to medicine have been derived accordingly. However, the life cycle of DCs, including cell origin, development, death and turnover, is only marginally understood. We employ computational methods to study the dynamics of the conventional DC (cDC) populations in mammalian immune systems. Recent experimental findings of dividing DCs in peripheral lymph organs require a mathematical model to illustrate the kinetics of DC homeostasis. To this end we build a steady-state model consisting of three linear ordinary differential equations (ODEs). Analytical solutions are derived directly from the model. And biological parameters of interest (including input rate, cell cycle length and death rate) are computed using data provided by our collaborator Dr. Kang Liu, which are consistent with other studies. In addition, stability of the model is confirmed by numerical simulations. We then extend the steady-state model to deal with stimulated states (equilibrium conditions altered by drugs, infections, or chemicals) by introducing the logistic growth model to cell proliferation with varying population sizes. The population of newly identified cDC precursors (pre-DCs) in spleen is included. In line with Dr. Liu's growth factor Flt3L-induced experiments, we conduct extensive computations to fine-tune the model with regions of parameters. We find the cell cycle is the most important vairable in response to the stress. Finally, we derive a systematic model for DC's dynamic life cycle, which unifies both steady- and stimulated-states. In the future, we want to apply our models to predict unforeseen changes, minor or drastic, induced by external stimulations in the DC and pre-DC populations. These computational undertakings is aimed to facilitate immunologists to effectively design their experiments to address key questions on pathogenic mechanisms in DC-related human disorders. The whole modeling approach is applicable to other biological systems and it can also be easily extended to systems of other cell types with knowledge of the developing pathways.

Bayesian analysis of gene-environment interactions with error in measurement of environmental covariates and missing genetic data

Iryna Lobach, Bani Mallick, and Raymond Carroll

Abstract: Case-control studies are widely used to detect gene-environment interactions in the etiology of complex diseases. Many variables that are of interest to biomedical researchers are difficult to measure at the individual level, e.g. nutrient intake, cigarette smoking exposure, long-term toxic exposure. Measurement error may cause bias in parameter estimates, thus masking key features of data. We develop a Bayesian methodology for analysis of case-control studies for the case when measurement error is present in an environmental covariate and the genetic variable has missing data. This approach offers several advantages. It allows prior information to enter the model to make estimation and inference more precise. The environmental covariates measured exactly are modeled completely nonparametrically. Further, information about the probability of disease can be incorporated in the estimation procedure to improve quality of parameter estimates. A unique feature of the procedure under investigation is that the analysis is based on a pseudo-likelihood function therefore conventional Bayesian techniques may not be technically correct. We propose an approach based on Markov Chain Monte Carlo sampling as well as a computationally simple method based on an asymptotic posterior distribution. Simulation experiments demonstrated that our method produced parameter estimates that are nearly unbiased even for small sample sizes. An application of our method is illustrated using a population-based case-control study of the association between calcium intake with the risk of colorectal adenoma development.

Assessing Value-Added Models (VAM) for teacher effectiveness

Cortley Logan and Kathleen Zarnitz
Stonehill College
Value-Added Models are generating a lot of interest as a means of measuring teacher effectiveness. We explore the power of a simple example of this type of model and its sensitivity to missing data.
Keywords: VAM power missing data

A Discussion of Uncongeniality for Synthetic Data
Bronwyn Loong, Carl M. Morris, and Donald B. Rubin
Department of Statistics, Harvard University
In the multiple imputation literature, uncongeniality refers to differences in types of input between the imputer, typically an agency, and the analyst. When releasing synthetic data in place of observed values to protect data confidentiality, the statistical agency and external analyst are always separate bodies, hence the potential for uncongeniality always exists. In this article, we present a definition of congeniality for multiple imputation for synthetic data. Our definition is motivated by common examples of uncongeniality, specifically ignorance of the original survey design in analysis of fully synthetic data, and situations when the imputation model and analysis procedure condition upon different sets of records. We conclude that our definition assists the imputer to identify the source of a discrepancy between observed and synthetic data analytic results. Motivated by our definition, we derive an alternative approach to synthetic data inference to recover the observed data set sampling distribution given the synthetic data.
Keywords: analysis procedure, data confidentiality, data utility, imputation model

Nutritional label use and causal effects on body mass index
Michael Lopez and Roee Gutman
Department of Biostatistics, Brown University
Disclosure of ingredients and inclusion of a standardized label has been required on all US food and beverage since the Nutritional Labeling and Education Act (1990). The comparison of health outcomes between label users and non- users, however, is difficult since these groups may differ with regard to socio-economic and demographic characteristics. Specifically, those who read nutrition labels are, on average, healthier. In assessing the effect of label use on BMI, our goal will be to mimic a randomized trial in which confounders are balanced and participants are assigned to varying levels of exposure, to create populations which differ only with respect to label use. The data stem from the 2005-06 National Health and Nutrition Examination Survey (NHANES), a combination of survey data which quantifies label use and physical examinations which measure BMI. In considering the causal effect of label use on BMI, a common statistical method to account for differences in self-selection of label use among participants is the propensity score. However, while traditional propensity score methods involve binary exposures, label use in the NHANES questionnaire is assessed using a five-point scale (never, rarely, sometimes, most of the time, always). In analyzing this data, Drichoutis et al. (2009) use several probit regression models for the propensity to use one level of label use versus another, for each of the 10 pairs of exposure levels, and suggest no strong causal effect of label use on BMI. Building on the ideas of Rosenbaum (1999), Imbens (2000), and Imai and Van Dyk (2004), we estimate an individual's nutritional label exposure using a generalized propensity score by fitting an ordered logit model. We create subclasses based on the linear predictor from the ordered logit fit, and confirm balance among our confounders by varying levels of label use within each subclass. Lastly, we estimate pooled effects of label use and find evidence of a decrease in BMI with increased exposure. We compare our results to those obtained by Drichoutis et al. (2009), and discuss the origin of the differences.
Keywords: propensity score, causal inference

Modeling Mortality and Longevity Risk
Mary M. Louie and Greta M. Ljung
AIR Worldwide Corporation

Unanticipated changes in mortality rates and longevity pose a risk to life insurers and pension funds. In particular, if longevity increases more than expected, annuity providers may be exposed to payout levels that are higher than what a company or fund originally planned for. Even a relatively small change in life expectancies could create solvency problems for pension plans and insurers. Financing mechanisms introduced to address this issue include the transfer of longevity risks to financial markets through longevity bonds and mortality-linked securities. Stochastic mortality models are needed for pricing of these instruments. We review and compare the forecasting performance of some alternative mortality models and describe a general framework for calibrating and improving forecasts.
Keywords: Stochastic mortality model, longevity, longevity bonds, risk transfer

On Fan's Adaptive Neyman Tests for Comparing Two Spectral Densities
Kewei Lu and Linyuan Li
Department of Mathematics and Statistics, University of New Hampshire
We consider tests for assessing whether two stationary and independent time series have the same spectral densities (or same auto-covariance functions). Both frequency domain and time domain test statistics for this purpose are reviewed. The adaptive Neyman tests are then introduced and their performances are investigated. Our tests are adaptive, i.e., they are constructed completely by the data and do not involve any unknown smoothing parameters. Simulation studies show that our proposed tests are at least comparable to current tests in most cases. Furthermore, our tests are much more powerful in some cases, such as against the long orders of autoregressive moving average (ARMA) models such as seasonal ARMA series.
Keywords: adaptive Neyman test, periodogram, seasonal ARMA model, spectral density

Sampling on Networks: A Model-Based Approach
Simon Lunagomez
Consider a social network and a variable of interest that can (potentially) be measured for the nodes of such network. For most applications, it is neither practical nor meaningful to observe the variable of interest (or response) for the whole network, instead, the most reasonable thing to do is to take a sample. We are interested in making inferences on a population quantity that is a function of both, the response variable and the network topology. We propose a general modeling strategy that takes into account all the relevant sources of uncertainty (at least from the point of view of the literature) and a Bayesian procedure able to deal with the generality implied by this new framework. A motivation for this work is given by the need for inferring the incidence of HIV in high-risk populations (e.g. men that have sex with men, intravenous drugs users) for the sake of health policy. We discuss how the our methodology should be understood within the framework proposed by Donald Rubin regarding the inference for a population quantity and the first steps to extend such framework so it allows the discussion of non-ignorable designs, which has been quite marginal within the Statistics literature.

Interim Sample Size Reassessment to Fixed Duration Trials
Joe Massaro (1) and Alison Pedley (2)
(1) Department of Mathematics and Statistics, Boston University, (2) Merck
The log-rank test is often performed in randomized clinical trials designed to assess the superiority of an experimental treatment over a control with respect to a time to event endpoint. Due to uncertainties of the effect size in the design stage, adaptive designs allowing for sample size adjustment at the time of interim analysis based on the conditional power of observing a significant result by the end of the trial are becoming increasingly popular. In this presentation, methodology for a 2-stage adaptive design based on the log-rank test is developed in the setting of fixed duration trials (trials where each patient is followed for the same fixed amount of time, regardless of study length) by adapting methodology originally developed by Li et al (2002, 2005) for maximum information trials (trials where each patient is followed until the end of the study, regardless of when patient was enrolled). Simulations were performed to evaluate the performance of the new methodology. By redefining the relationship between the observed

number of events and the final critical value and removing the restriction on the maximum value of the interim efficacy boundary in the methodology of Li et al, the methodology developed here achieves the desired interim conditional power while still maintaining control of the type I error rate.

PANEL: Teaching Business Statistics: Best Wisdom and Biggest Challenges
John McKenzie, Tevfik Aktekin, Keith Ord, and Erol Pekoz
This panel discussion will cover issues relating to teaching business statistics. Each panelist will briefly speak about the best wisdom they have and the biggest challenges they face about teaching business statistics, and then the floor will open for discussion and questions from the audience. Audience members and panelists are encouraged to create a lively discussion by sharing the best examples, stories, exercises, datasets, games, etc. that they use in class. The panelists have all been very successful teachers; this session is also appropriate for people who do not teach business statistics, but who instead teach other types of introductory statistics courses.

Considerations for a Career as a Statistician in the Pharmaceutical Industry
Ronald Menton
Purchasing the first home is a daunting task for many new home owners due to the numerous options available to them. Potential new home buyers are encouraged to talk to friends, family, and other experts to help them develop and understand the considerations important to them for their house purchase decision process. Similarly, selecting the first job can be a daunting task for the newly graduated statistician. This talk will apply the analogy between a house purchase and selecting the first career to illustrate important factors for considering a career as a statistician in the Pharmaceutical Industry. The presentation will summarize the various careers for a pharmaceutical statistician, discuss opportunities for career growth, share some observations on the statistical community, and outline some trends for the career as a pharmaceutical statistician.

Bayesian framework for the incorporation of multiple data sources with the analysis of infectious disease outbreaks
Carlee B. Moser (1), Mayetri Gupta (1), Brett N. Archer (2), and Laura F. White (1)
(1) Department of Biostatistics, Boston University School of Public Health, (2) National Institute for Communicable Diseases (NICD), National Health Laboratory Service (NHLS), South Africa
When an outbreak of an infectious disease occurs, public health officials need to understand the dynamics of disease transmission in order to launch an effective response. Two quantities that are necessary to describe transmission are the reproductive number and the distribution of the serial interval. The basic reproductive number, R_0, is the average number of secondary cases a primary case will infect, assuming a completely susceptible population, as might exist initially in the outbreak of a novel disease. The serial interval (SI) provides a measure of temporality, and is defined as the time between symptom onset between a primary case and its secondary case, and is typically summarized by the mean of this distribution, μ. We present a Bayesian framework to estimate the reproductive number and serial interval using the method introduced by White and Pagano (2008), who proposed a maximum likelihood technique for the simultaneous estimation of R_0 and SI using the epidemic curve. This Bayesian framework allows for the incorporation of additional data sources beyond the epidemic curve through the prior distributions. We explore the inclusion of a contact trace sample or household study data through simulation, and apply these methods to outbreak data from the SARS outbreak in Hong Kong and Singapore in 2003, and the influenza A(H1N1)2009pdm outbreak in South Africa. The use of prior information decreases the bias and variability of the estimates. When no additional data is used to inform the prior, the Bayesian estimates are less variable than those obtained by the original White and Pagano method. The results from the South African H1N1 analysis illustrate this, with $R_0 = 1.462$ (95% Credible Interval: 1.337–1.597) when no contact trace data is used, and $R_0 = 1.409$ (95% CI: 1.32–1.509) when contract trace is included. The White and Pagano estimate is comparable, but has larger variability with $R_0 = 1.465$ (95% CI: 1.08–2.81). The serial interval results follow similarly with $\mu = 2.66$ (95% CI: 2.031–3.286)

when no contact trace data is used, and $\mu = 2.326$ (95% CI: 1.921–2.81) when the prior is informed with the contract trace data. The White and Pagano estimate is slightly larger for μ and has larger variability with $\mu = 2.699$ (95% CI: 1.354–4.374). This Bayesian approach we have introduced allows for a more flexible and stable estimation framework with improved diagnostic capabilities. Prior information can be included which greatly improves estimation of key transmission parameters, by reducing both the bias and variability of the estimates.
Keywords: Bayesian, infectious disease, reproductive number, serial interval

Model-Based Estimation for Respondent-Driven Sampling
Sergiy Nesterko
Respondent-Driven Sampling (RDS) is a standard process for sampling from hard-to-reach populations such as injection drug users or men who have sex with men, widely employed by public health agencies around the world. RDS recruits respondents by tracing links in the network of connections of individuals from the underlying hidden population, with current participants financially encouraged to refer next ones. While the process is very effective in gaining access to hidden populations, it introduces complex dependence patterns in collected information and poses statistical challenges when estimating population quantities, such as population averages. Here we develop the first model-based procedure for RDS to estimate population averages, as well as its extensions to handle missing data and different participant types, such as "close" or "distant" friends etc. The approach does not model the underlying network of individuals, and models the dependence in observations by assuming homophily, i.e. the tendency of individuals with similar measurements to connect more often. Such approach results in wider uncertainty intervals than produced by existing techniques. Via simulation, we show the method's superiority in terms of uncertainty intervals coverage rates and mean squared error (MSE) when comparing to the current estimation techniques. We demonstrate an application of our method to a survey of individuals at high risk of HIV conducted in San Diego,and suggest that it may lead to more accurate estimation for RDS-based surveys.

Modeling Neural Spiking Activity in Patients with Parkinson's Disease During Movement and at Rest
Thien Tai T. Nguyen and Uri T. Eden
Department of Mathematics and Statistics, Boston University, Undergraduate Research Opportunity Program (UROP)
Brain areas communicate and process information through sequences of electrical impulses, called "spikes." In a healthy brain, sequences of spikes provide vital information about how we react to stimuli and interact with the world around us. In brains with neurological disorders, such as Parkinson's disease, there are alterations in the spiking activity that interfere with the informational content of these signals. Deep brain stimulation is a surgical treatment for Parkinson's disease that alleviates tremor and helps patients initiate movements by constantly stimulating altered brain areas with electric impulses. Electrical recordings of brain activity are collected as part of the surgery to implant the device. These recordings provide a unique opportunity to explore brain function during movement planning and execution. Data was collected from patients during a surgical procedure to implant a deep brain stimulator. The patients performed a visually guided movement task by using a joystick to move an onscreen cursor to one of four targets. They repeated this task over multiple trials, while neural impulses from the STN were recorded. We examined the spiking data for each trial from a period 1.6 seconds before the start of the joystick movement to 1.6 seconds after the start of movement. We analyzed the data to describe the properties of the spiking rate, distribution of spiking, and dependence of current spiking on past spiking activity during this task. We analyzed this dependence structure in the time domain using auto-correlation plots, and in the frequency domain by computing estimates of the spectrum of the spiking activity for each neuron. In addition, we analyzed the relationship between the firing activity of one cell with the activity of other simultaneously recorded cells using cross-correlation plots in the time domain, and coherence and phase plots in the frequency domain. Finally we constructed generalized linear models to describe

structure of spiking in relation to movement variables and history dependence. We analyzed a total of 25 cells, and made comparisons between 19 pairs of simultaneously recorded cells, recorded over 13 different experimental sessions. We computed auto-correlation plots for each cell, and found significant correlations at small time lags (between ±100 milliseconds) during the period before movement that were not present during movement. Our spectral estimators indicated that before movement, many neurons fired rhythmically with frequencies around 10-20 Hz, while during movement, these spiking rhythms vanished. We also found that pairs of cells would fire coherently before movement, but not after movement. GLMs incorporating history has shown significant history dependence structures present in the before stage but not in the after stage. One theory of spiking activity in this region holds that aberrant oscillatory activity prevents motor signals from being properly transmitted to downstream brain areas. Our findings support this theory. Before movement, when the patients are more likely to have difficulty moving, we found large peaks in the power spectra at about 10-20Hz, and strong coherence at these frequencies between cells. This coherent oscillatory activity disappears once the patient is able to start the movement. This suggests that breaking the pattern of oscillations in the STN to allow activity similar to natural movement might be one mechanism by which DBS works. Further study of these neural spiking properties may help us develop new and more effective therapies.
Keywords: neurology neuroscience spiking activity neurons GLM generalized linear models Parkinson's disease

Network Mega Analysis: Application to Historically Controlled Clinical Trials
A James O'Malley
Department of Health Care Policy, Harvard Medical School
I describe a novel design for a single armed trial of a new drug coated coronary-artery stent. In order to make comparisons against coronary artery bypass graft (CABG) surgery, two historical trials involving a third treatment (bare metal stenting) are used. Because one of the historical trials is a small study comparing the drug coated stent to a bare metal stent and the other is a large randomized trial comparing the same bare metal stent to CABG, the pairs of treatments form a connected network between the trials. I will discuss the virtues of the network meta-analysis design when individual level data are available and will present the results from the actual analysis of the trial. In particular, I will demonstrate that a different conclusion is obtained than for the standard historical-control analysis.
Keywords: Bayesian analysis, Coronary-artery stents, Historical control, Hierarchical modeling, Medical devices

Generalized Kiefer Process and Queues
Guodong Pang (1) and Ward Whitt (2)
(1) Penn State University, (2) Columbia University
In this talk, we will discuss generalized Kiefer process and its use in the study of the impact of dependent service times in queues with many servers. We show that under certain mixing conditions on the sequence of successive service times, the number of busy servers in the infinite server queues can be approximated by a Gaussian process (random field) driven by a generalized Kiefer process. We characterize the effect of the dependence among service times upon the mean and variance in steady state. We also use this result to approximate the delay probability in many-server queues with dependent service times.

Bayesian Degree-corrected Stochastic Block Models for Community Detection
Lijun Peng and Luis E. Carvalho
Department of Mathematics and Statistics, Boston University
We discuss a degree-corrected version of a stochastic block model that aims to achieve a better resolution for community identification. We follow a fully Bayesian approach and conduct inference based on a principled centroid estimator of community labels. To this end, an efficient Gibbs sampler is developed. We demonstrate the proposed model and inference on a classical network dataset. Finally, we offer a few concluding remarks on the model implementation and directions for future work.

Keywords: Bayesian, network detection

Insufficiency and the Preservation of Fisher Information

David Pollard
Yale University

A discussion of some statistical ideas related to a simple example, due to Kagan and Shepp, showing that Fisher information can be preserved by statistics that are not sufficient.

Keywords: Fisher information; Hellinger di

Profile likelihood based confidence interval for the difference between two correlated proportions

Vivek Pradhan, Krishna K Saha, Tathagata Banerjee, Yuanye Zhang, and John C Evans
1. Boston Scientific Corporation, 100 Boston Scientific Way, Marlborough, MA 01752, USA 2. Department of Mathematical Sciences, Central Connecticut State University,New Britain, CT 06050, USA 3. Indian Institute of Management, Ahmadabad, India 4. Department of Statistics, University of Connecticut, CT, USA

Inference concerning the difference between two binomial proportions in the paired data is often an important problem in many biomedical investigations. Based on the method of variance of estimates recovery (MOVER) Tang et al. (2010, Statistics in Medicine) discussed six methods to construct confidence intervals for the difference between two correlated proportions. In this article, using the different choices of adjustments to the cell frequencies of a 2×2 table as proposed by Agresti and Min (2005, Statistics in Medicine), we propose three new approaches based on profile likelihood method. Through simulations, in terms of coverage probabilities and expected lengths, we then compared our proposed methods with the methods recommended by Tang et al. (2010). Simulation study indicates that our proposed approaches perform reasonable well; cover probabilities are closer to the nominal level, and expected interval lengths are competitive to Tang et al. Finally, we illustrate the proposed confidence intervals with two real-life examples.

Keywords: confidence interval; correlated binomial proportions; Jeffreys prior; paired data; profile likelihood

Finite Mixture Models for Biomedicine at the Single Cell Level

Saumyadipta Pyne
Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School

Biomedicine at high-resolution single-cell level is made possible by reliable characterization of the cellular state-spaces defined by complex biological networks such as in cell signaling or differentiation. High-dimensional multi-parametric flow cytometry could be used to generate per cell data for investigating such networks. To address the unique challenges of such data, we constructed several new platforms for automated single cell multi-marker expression data analysis. We introduced new statistical methodology in the form of finite mixture models of multivariate skew t and skew normal distributions that are fit with new and efficient Expectation-Maximization algorithms. Such parametric models allow robust modeling against asymmetry and outliers in data. Thus the algorithms model cell populations with complex shapes and structures within each sample in high-dimensional marker-space, and then employ graph matching for registration of the corresponding cell populations across samples to enable comparison of different phenotypes and time points. The platforms have opened up new possibilities for many biomedical applications, some of which will be illustrated with examples from stem cell systems biology, cell signaling, and clinical studies.

Keywords: Finite mixture model, Flow cytometry, Clustering

Analysis of Catch Counts with Excess Zeros and Over-Dispersion in Bayesian Approach

Rengui Qiao and Liliana Gonzalez
Department of Computer Science and Statistics, University of Rhode Island

In fisheries, catch counts consist of non-negative values, usually with excess zeros and a heavy right tail. Several generalized linear models can be used in the analysis of such data since these methods do not rely on the assumption of normality. The aims of our study include two parts, to reveal appropriate models in fitting the black sea bass catch count data and to evaluate the effects of vent sizes on the catch efficiency, after removing important effects such as vessel, string within vessel, soaking time, and fishing pot position. In the experiments, fishing pots were investigated with various circular escape vent sizes (0", 2.38", 2.75", 3.10", and 3.40" as the diameter). Our main interest is to reveal the ideal vent size which gives the best legal size black sea bass catch count and minimizes the sublegal size catch count. The Poisson regression model is the most widely used methodology to analyze count data. Different Poisson models are used in the analysis of the black sea bass catch count. We start with the standard Poisson model, and continue with the regression for Poisson rates. In order to account for over-dispersion, a Poisson model with scaled deviance as well as a negative binomial regression model are discussed. Since the count data contain excess zero counts, a Zero Inflated Poisson (ZIP) model is used. From the results of the last analysis, we can conclude that over-dispersion is not only induced by excess zero counts in the data, but also contributed by the right tail. Because of the hierarchical structure in parts of the data, random components are introduced in the above mentioned models by assuming random intercepts on the subjects. The mixed effect models are fitted to the black sea bass catch count data using Poisson, negative binomial, ZIP and Zero Inflated Negative Binomial (ZINB) models. All mixed effect models are fitted using Bayesian approach in Winbugs. After comparison of the performance of all models, the ZINB model is recognized as the best fitting model as indicated by smallest Deviance Information Criterion (DIC) and Negative Cross-validatory Log-likelihood (NLL) values. Ninety-five percent credible intervals are computed for both legal and sublegal size black sea bass, and we conclude that vent sizes of 2.38" and 2.75" have higher average counts of legal size black sea bass than those of 3.10" or 3.40", after adjusting for important covariates; vent sizes of 2.75" have significantly less sublegal catch counts than vent sizes of 2.38" and 3.10", as well as having a tendency of less sublegal catch than 3.40". Therefore, we can conclude that a vent size of 2.75" is the ideal choice, since it maximizes the retention of legal size black sea bass and at the same time minimizes the retention of sublegal size black sea bass in the fishing pots.

Keywords: catch count, hierarchical structure, over-dispersion, zero-inflation, Poisson, NB, ZIP, ZINB, bayesian

Optimal Stopping Time for the Last Passage Time and Maximum Time

Dan Ren and Kostas Kardaras
Department of Mathematics and Statistics, Boston University

Given a transient diffusion process X, let ρ_p be the last time when X passes level ℓ and ρ_m be the last time when X reaches the maximum. For each random time ρ_p and ρ_m, this paper solves the optimization problem $\inf_\tau E[\lambda(\tau-\rho)_+ + (1-\lambda)(\rho-\tau)_+]$ over all stopping times τ of X. The optimal stopping time of ρ_p is given as $\tau_p = \inf\{t \in R^+ : X_t \leq \kappa\}$ where κ is the solution of an explicit defined equation. The optimal stopping time of ρ_m is given as $\tau_m = \inf\{t \in R^+ : X_t \leq f(\sup X_s), 0 \leq s \leq t\}$ where the function f is the maximal solution of a first-order ordinary differential equation

Testing for the homogeneity of the dispersions in the clustered count data

Krishna K. Saha
Department of Mathematical Sciences, Central Connecticut State University

This article develops procedures for the validity of the equal dispersion assumption for testing the equality of the means among treatment groups in the analysis of clustered count data. The $C(\alpha)$ tests based on likelihood and quasi-likelihoods are obtained. Monte Carlo simulations are then used to study the comparative behavior of these $C(\alpha)$ statistics in terms of size and power. The simulation results demonstrate that the $C(\alpha)$ statistics based on quasi-likelihoods hold the nominal level reasonably well. Finally, an application to toxicological data is presented.

Confidence Intervals for the Dispersion Parameter in the Clustered Count Data

Debaraj Sen (1), Krishna K. Saha, and Chun Jin (2)
(1) Department of Mathematics and Statistics, Concordia University, (2) Department of Mathematical Sciences, Central Connecticut State University
It is challenging to make inferences about the dispersion parameter in the clustered count data occurring in biomedical studies. A couple of procedures for the construction of confidence intervals of the dispersion parameter have been investigated, but little attention has been paid to the accuracy of its confidence interval. In this article, we consider various approaches for computing the confidence intervals in counts based on the parametric and semiparametric models. Numerical studies including simulations and real data examples are presented.
Keywords: asymptotic confidence interval; dispersion parameter; hybrid profile variance approach; profile likelihood approach; non-parametric bootstrap approach

A Functional Method for Conditional Logistic Regression with Errors-in-Covariates
Samiran Sinha
Texas A and M University
In this talk we describe a functional approach for handling errors in covariates in matched case-control studies which are commonly analyzed through the conditional logistic regression. We propose to estimate the parameters from a set of unbiased estimating equations. The proposed method requires that the moment generating function of the measurement errors exists. We also investigate the asymptotic properties of the estimators. The finite sample performance of the method is judged via simulation studies. The proposed methodology is illustrated by analyzing the data from the NIH-AARP Diet and Health study.

Omnibus Risk Assessment via Accelerated Failure Time Kernel Machine Modeling
Jennifer A. Sinnott and Tianxi Cai
Department of Biostatistics, Harvard School of Public Health
Integrating genomic information with traditional clinical risk factors to improve the prediction of disease outcomes could profoundly change the practice of medicine. However, the large number of potential markers and the complexity of the relationship between markers and disease make it difficult to construct accurate risk prediction models. Standard approaches for identifying important markers often rely on marginal associations and may not capture non-linear or interactive effects. At the same time, much work has been done to group genes into pathways and networks. Integrating such biological knowledge into statistical learning could potentially improve model interpretability and reliability. One effective approach is to employ a kernel machine (KM) framework, which has been recently extended to analyzing survival outcomes under the Cox model. In this paper, we propose KM regression under the accelerated failure time model. We derive a pseudo score statistic for testing and a risk score for prediction of survival. To approximate the null distribution of our test statistic, we propose resampling procedures which also enable us to develop alternative robust testing procedures that combine information across kernels. Numerical studies show that the testing and estimation procedures perform well. The methods are illustrated with an application in breast cancer.
Keywords: Accelerated Failure Time Model, Kernel Machine, Resampling, Risk Prediction, Survival Analysis

Aspects of Presidential Voting (and Implications for 2012)
Robert B. Smith
Social Structural Research Inc.
Providing a theoretical and empirical paradigm for the study of presidential elections, this presentation provides an in-depth analysis of a major election-night telephone survey of voters in the 2008 election. It focuses on the impacts of variables of social structure, ideology, party affiliation, the issues and facilitating factors. The interrelationships among these factors are modeled using multilevel and structural equation statistical procedures. Some surprising findings: Obama and McCain were about even among

traditional voters, but early deciders, early voters, eager voters, mail-in voters, those who only voted for the presidential candidates, and those who did not vote in the 2006 congressional election were more likely to vote for Obama. In terms of ideology, political moderates were more likely to support Obama than McCain. The mechanism linking political ideology with party affiliation and then to vote is strongly supported in these data. When these intervening variables are controlled, the Red-Purple-Blue typology of states and a measure of economic distress have little direct effects on vote in these multilevel models; there effects are largely indirect. The summarizing graphical models of the relationships among these pivotal factors are estimated using structural equation procedures. The implications of these findings for 21st century electoral politics are briefly discussed.
Keywords: Presidential Elections

Variable Selection for High-Dimensional Multivariate Outcomes with Application to Genetic Pathway/Network Analysis
Tamar Sofer (1), Lee Dicker (2), and Xihong Lin (1)
(1) Department of Biostatistics, Harvard School of Public Health, (2) Department of Statistics, Rutgers University.
We consider variable selection for high-dimensional multivariate regression using penalized likelihoods when the number of outcomes and the number of covariates might be large. To account for within-subject correlation, we consider variable selection when a working precision matrix is used and when the precision matrix is jointly estimated using a two-stage procedure. We show that under suitable regularity conditions, penalized regression coefficient estimators are consistent for model selection for an arbitrary working precision matrix, and have the oracle properties and are efficient when the true precision matrix is used or when it is consistently estimated using sparse regression. We develop an efficient computation procedure for estimating regression coefficients using the coordinate descent algorithm in conjunction with sparse precision matrix estimation using the graphical LASSO (GLASSO) algorithm. We develop the Bayesian Information Criterion (BIC) for estimating the tuning parameter and show that BIC is consistent for model selection. We evaluate finite sample performance for the proposed method using simulation studies and illustrate its application using the type II diabetes gene expression pathway data.
Keywords: Efficiency; Gene set analysis; Joint estimation; Model selection; Multiple outcomes; Sparsity.

Absolute protein quantitation: Inference with non-ignorable missing data in high throughput proteomics
Eric Solis (1,2) and Edoardo Airoldi (2)
(1) Program in Systems Biology, Harvard University, (2) Harvard University, Department of Statistics; Harvard FAS Center for Systems Biology; The Broad Institute of MIT and Harvard
Recent advances in high-throughput proteomics enable the detection of thousands of proteins in complex biological samples through the use of mass spectrometry. However, technical factors obviate the detection of every protein and and the identification process itself is biased towards high-abundance proteins. These limitations induce complex patterns of missing data where missingness and abundance are anticorrelated, which prevents the quantitative information corresponding to identified protein from being used to determine the absolute abundance of these proteins. In this work we develop a Bayesian statistical model and inference algorithm to estimate parameters in a hierarchical model from data with a large, but unknown amount, of missing data generated by multiple missingness mechanisms. Further, we exploit analytic computations in combination with efficient numerical routines to efficiently impute the non-ignorable missing data, avoiding the Random Walk Metropolis Hastings Algorithm canonically used to sample from variable-dimension posterior distributions. We demonstrate via simulation and experimental data that our method out performs existing quantitative strategies and allows the accurate determination of protein abundance with dynamic range equivalent to that of cellular proteins.
Keywords: proteomics; mass spectrometry; non-ignorable missing data; imputation; exact sampling methods

Graphlet decomposition of a weighted network

Hossein Azari Soufiani and Edoardo M. Airoldi
Department of Statistics, Harvard University
We introduce the graphlet decomposition of a weighted network, which encodes a notion of social information based on social structure. We develop a scalable inference algorithm, which combines EM with Bron-Kerbosch in a novel fashion, for estimating the parameters of the model underlying graphlets using one network sample. We explore some theoretical properties of the graphlet decomposition, including computational complexity, redundancy and expected accuracy. We demonstrate graphlets on synthetic and real data. We analyze messaging patterns on Facebook and criminal associations in the 19th century.
Keywords: Expectation-Maximization; Bron-Kerbosch; sparsity; deconvolution; massive data; statistical network analysis; parallel computation; social information.

Uncertainty Propagation from Network Inference to Characterization
Weston Viles, Prakash Balachandran, and Eric D. Kolaczyk
Boston University
Network-based data (e.g., from sensor, social, biological, and information networks) now play an important role across the sciences. Frequently the graphs used to represent networks are inferred from data. Surprisingly, however, in characterizing the higher-level properties of these networks (e.g., density, clustering, centrality), the uncertainty in their inferred topology typically is ignored. The distribution of estimators characterizing these networks defined implicitly through standard thresholding procedures can have distributions complicated by dependence inherent among the thresholded events. Motivated by this observation, we present a method by which the distribution of a sum of dependent binary random variables is approximated and demonstrate the method by exploring the problem of estimating network density - a simple but fundamental characterization of a network - in the context of correlation networks with Gaussian noise.

Dynamic Portfolio Execution
Chun Wang
Department of IE and OR, Columbia University
We consider the problem of dynamically purchasing or selling a fixed portfolio of securities when there is a price impact associated with trading. These problems suffer from the curse of dimensionality and so we seek good sub-optimal strategies using approximate dynamic programming and other heuristic techniques. We consider problems with no short-sales and other portfolio constraints and use recently developed duality techniques to evaluate the quality of our policies.

Model Selection for Cox Models with Time-varying Coefficients
Jun Yan (1) and Jian Huang (2)
(1) University of Connecticut, (2) University of Iowa
Cox models with time-varying coefficients offer great flexibility in capturing the temporal dynamics of covariate effects on right censored failure times. Since not all covariate coefficients are time-varying, model selection for such models presents an additional challenge, which is to distinguish covariates with time-varying coefficient from those with time-independent coefficient. We propose an adaptive group lasso method that not only selects important variables but also selects between time-independent and time-varying specifications of their presence in the model. Each covariate effect is partitioned into a time-independent part and a time-varying part, the latter of which is characterized by a group of coefficients of basis splines without intercept. Model selection and estimation are carried out through a fast, iterative group shooting algorithm. Our approach is shown to have good properties in a simulation study that mimics realistic situations with up to 20 variables. A real example illustrates the utility of the method.

Estimating Network Degree Distributions from Sampled Networks: An Inverse Problem
Yaonan Zhang, Eric D. Kolaczyk (1), and Bruce D. Spencer (2)
(1) Department of Mathematics and Statistics, Boston University, (2) Department of Statistics, 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. We study the problem of how to estimate the degree distribution of a true underlying network from its sampled network, focusing on the case of induced sub-graph sampling. We show that it can be formulated as an ill-posed inverse problem. Accordingly, we offer a penalized least-squares approach to solving this problem, with the option of additional constraints. The resulting estimator is a linear combination of singular vectors of a matrix, relating the expectation of our sampled degree distribution to the true underlying degree distribution, which is defined entirely in terms of the sampling plan. We apply our method to simulated and real data for both homogeneous and inhomogeneous networks. Our results show that the estimators from both types of networks can be appropriately smoothed with proper choice of the penalization parameters. We explore various approaches to selecting the penalization parameter, including methods based on the bootstrap and Stein's unbiased risk estimation.
Keywords: Networks, degree distribution, inverse problem, penalized least-squares.

Optimalities in Estimation of Large Precision Matrices
Harrison Zhou
Department of Statistics, Yale University
In this talk I will present some preliminary results on optimal estimation of large precision matrices. Two settings will be considered: sparse and latent-variable graphical models.

A Novel Method to Include Phenotyped but Ungenotyped Relatives in Genetic Association Tests
Wei Vivian Zhuang (1), Ching-Ti Liu (1), Gyungah Jun (1), Kathryn L. Lunetta (1,2,3)
(1) Department of Biostatistics, Boston University School of Public Health, (2) Sections of General Internal Medicine, Preventive Medicine and Epidemiology, Department of Medicine, Boston University School of Medicine, (3) National Heart, Lung and Blood Institute Framingham Heart Study
In some long-term longitudinal studies such as the Framingham Heart Study, there are individuals with rich phenotype data who died before providing DNA for genetic studies. Thus, the individuals have no genotype data but have phenotypic data. Often, the genotypic and phenotypic data of the relatives are available. Visscher and Duffy (2006) and Chen and Abecasis (2007) explored the power increase due to the inclusion of ungenotyped individuals in a genetic association test for a quantitative trait. Both studies inferred missing genotypes based on observed genotypes. We propose a phenotypically enriched genotypic imputation (PEGI) method to impute missing genotypes using observed phenotypes in addition to genotypes. Our simulations with genotypes missing completely at random (MCAR) show that, for a SNP with moderate to strong effect on a phenotype, PEGI improves power more than imputation based solely on genotypes without excess type I errors. The effect estimate is often biased when the outcome is used for imputation while it is unbiased when a phenotype unrelated with the outcome is used. Compared to using only the observed genotypes for imputation, the PEGI method may improve power for data with genotypes MCAR, missing at random (MAR), or not missing at random (NMAR).
Keywords: Genotype Imputation, Missing Genotypes, and Family Studies.

New statistical approaches to missing heritability
Or Zuk (1), Eliana Hechter (1), Shamil Sunyaev (1,2), and Eric Lander (1)
(1) Broad Institute of MIT and Harvard, (2) Brigham and Women's Hospital, Harvard Medical School
Genome-Wide Association Studies (GWAS) have been successful in discovering thousands of statistically significant, reproducible, genotype-phenotype associations in humans. However, the discovered variants (genotypes) explain only a small fraction of the phenotypic variance in the population for most human traits. In contrast, the heritability, defined as the proportion of phenotypic variance explained by all genetic factors, was estimated to be much larger for those same traits using indirect population-based estimators. This gap is referred to as *missing heritability*. Mathematically, heritability is defined by considering a function F mapping a set of (Boolean) variables, $(x_1, .., x_n)$ representing genotypes, and

additional environmental or *noise* variables, to a single (real or discrete) variable z, representing phenotype. We use the variance decomposition of F, separating the linear term, corresponding to additive (narrow-sense) heritability, and higher-order terms, representing genetic-interactions (epistasis), to explore several explanations for the missing heritability mystery. We show that genetic interactions can significantly bias upwards current population-based heritability estimators, creating a false impression of missing heritability. We offer a solution to this problem by providing a novel consistent estimator based on unrelated individuals. We also propose novel estimators for the different variance components (beyond additive) of heritability from GWAS data. Finally, we use the Wright-Fisher process from population genetic theory to study the relative contributions of rare and common variants to heritability.
Keywords: statistical genetics, missing heritability, epistasis, variance components estimation

Instruction for Wireless Internet Use

1. Wireless network name: bmc-guest
2. Open browser, and enter your email address when prompted to register

Figure 1: Boston University Medical Campus