QPRC 2017

The 34th Quality and Productivity Research Conference
Quality and Statistics: A Path to Better Life

June 13-15, 2017
The University of Connecticut, Storrs

QPRC 2017 Program

Contents

UCONN
UNIVERSITY OF CONNECTICUT
College of Liberal Arts & Sciences

University of Connecticut
Office of the Dean

June 2017

Dear QPRC scholars and professionals,

Welcome! We are pleased to host the 34th Quality and Productivity Research Conference. I understand that this year's attendee list boasts more than 130 participants, from across the nation plus several visitors from overseas. We are especially happy to have almost 50 students in attendance this year, and are grateful for the generous funding from the National Science Foundation, the Natrella scholarship, and the QPRC scholarship. I know that you will all benefit from engaging with your colleagues as you share research and discuss emerging issues in your field.

Since its founding in 1939, the College of Liberal Arts and Sciences has been the academic heart of the University of Connecticut. We take seriously the foundations of a liberal education: We teach students to think creatively and analytically; to reason from evidence; to respect the views and experiences of all members of our diverse community; and to continue learning throughout their lives, wherever their professional and personal journeys take them. It is wonderful to hear that the QPRC 2017 upholds these same values.

Best wishes,

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

An Equal Opportunity Employer

215 Glenbrook Road, U-4158
Storrs, Connecticut 06269-4158

Telephone: (860) 486-1231
Facsimile: (860) 486-0296
Web: http://bsc.clas.uconn.edu

University of Connecticut
Department of Statistics

College of Liberal Arts
and Sciences

A Message from the Department Head

It is my great pleasure to welcome all of you to attend QPRC: The 34th Quality and Productivity Research Conference. On behalf of the Department of Statistics at the University of Connecticut, I would like to wish all the delegates having an enjoyable visit to our Storrs campus as well as the Storrs center. We are thrilled to have this year three distinguished plenary speakers: Professor Shelemyahu Zacks of Binghamton University, Professor William Q. Meeker of Iowa State University, and Dr. Vijay Nair of Wells Fargo. I would also like to thank Dr. Steven L. Scott of Google for offering a short course on "Computational Bayesian Methods for Big Data Problems" on Monday, June 12, 2017.

My sincere thanks go to the leadership of Professors Nalini Ravishanker (Chair) and Haim Bar (Co-Chair) for organizing this conference. I would like to congratulate the program committee for putting together a wonderful conference program. I further thank Ms. Tracy Burke, Henry Linder, and student volunteers for their help at each stage of the planning for this conference. Finally, I would like to take this opportunity to acknowledge our generous sponsors: College of Liberal Arts & Sciences Dean's Office, Minitab, NSF, SAS / JMP, Stat-Ease, TriloByte, and UConn Department of Statistics.

Welcome to UConn and Storrs.

With best wishes,
Ming-Hui Chen, Professor and Head

Conference Honoree

Professor Shelemyahu Zacks

Binghamton University

Professor Shelley Zacks has an international reputation for his ground-breaking work in several areas that are directly connected with Quality and Productivity including applied probability, Bayesian sequential inference, change-point problems, common mean problems, experimental designs, geometrical probabilities, queueing systems and inventory, reliability and life testing, sequential methodologies, stochastic processes, survey sampling, actuarial studies, tracking, and filtering. A lifelong passion for handling difficult problems arising in science and engineering has been a primary inspiration behind most of his research and about 250 publications. Prior to 1980, his career path took him to the Technion, NYU, Stanford University, Kansas State University, University of New Mexico, Tel Aviv University, Case Western Reserve, and Virginia Tech. At SUNY-Binghamton (now Binghamton University) which he joined in 1980, he was Professor, Chair of the Department of Mathematical Sciences, and Director of the Center for Statistics, Quality Control, and Design. Professor Zacks has served as Joint Chief Editor, a Coordinating Editor, an Advisory Editor, and the Executive Editor for the *Journal of Statistical Planning and Inference*. He has also served on the editorial board of several journals including *Journal of the American Statistical Association, Annals of Statistics, Journal of Statistical Planning and Inference, Naval Research Logistics Quarterly, Communications in Statistics, American Journal of Mathematical and Management Sciences, Methodology and Computing in Applied Probability*, and *Sequential Analysis*.

Professor Zacks has earned many honors and awards, including Fellow of the Institute of Mathematical Statistics (1973), Fellow of the American Statistical Association (1974), Fellow of the American Association for the Advancement of Science (1983), elected membership in the International Statistical Institute (1975), honorary degree of Doctor of Philosophy from the University of Haifa (2005), and the Abraham Wald Prize in Sequential Analysis (2008).

In 1957 Professor Zacks published his first paper on wind produced energy and its relation to wind regime. Since then he has published numerous articles in refereed journals, twenty seven chapters in books, eighteen technical reports and several book reviews. His recent paper in the prestigious *Journal of Applied Probability* is on "Compound Poisson Processes With Poisson Subordinator". In addition, Professor Zacks has published nine books. During his academic career he has ably guided thirty Ph.D. students, and supervised four Master theses.

The primary research papers of Professor Zacks can be classified into the following fourteen areas of mathematical statistics, applied probability, and applications to various areas of quality and productivity.

1. Estimation, Testing and Decision Theory: 21 papers and 3 book chapters
2. Design of Experiments: 8 papers and a book chapter
3. Finite Populations Analysis: 10 papers and a book chapter
4. Change-Points Analysis: 8 papers and 3 book chapters
5. Adaptive Procedures: 5 papers and 2 book chapters
6. Sequential Estimation: 10 papers and a book chapter
7. Distribution of Stopping Times: 20 papers and a book chapter
8. Reliability Analysis and Quality Control: 5 papers and 8 book chapters
9. Inventory Systems: 13 papers and a book chapter
10. Queuing Systems: 7 papers and a book chapter
11. Stochastic Visibility: 5 papers and a book chapter

12. Actuarial Science: 2 papers and a book chapter
13. Applications in Physics, Biology, Genetics: 15 papers
14. Military OR: 4 papers and a book chapter.

His papers on Distribution of Stopping Times (Area 7) developed new methods for the exact evaluation of these distributions. The papers on Sequential estimation (Area 6) apply the methods developed in Area 7 to determine exact computations of the characteristics of the procedures. Most of the papers in the area of Design of Experiments (Area 2) developed a new methodology of randomized fractional replications, i.e., to choose a block of trials at random in order to eliminate the bias due to significant aliases. The papers on Finite Populations Analysis (Area 3) deal both with design and modeling approaches. The first paper cited on Change-Points Analysis (Area 4) is the seminal paper of Chernoff and Zacks. This paper developed a new test statistic for the existence of a change point, and has been cited in almost all papers written afterwards. All the papers on Adaptive Procedures (Area 5) developed a new approach for the search of the maximum tolerated dose (MTD) in Cancer Phase I Clinical Trials. Professor Zacks is world renowned and has been invited during the years to numerous international conferences to organize sessions and to deliver invited talks. Currently, he is Emeritus Distinguished Professor at Binghamton University, having retired in 2014.

Degrees

- B.A. Hebrew University 1955
- M.Sc. Technion (Israel) 1960
- Ph.D. Columbia University 1962

Plenary Talk - Professor Shelemyahu Zacks

Title: The Operating Characteristics of Sequential Procedures in Reliability
Chair: **Emmanuel Yashchin**, IBM T. J. Watson Research Center

Abstract: Sequential methods in reliability testing have been in use since the early 50s. Sequential life testing and sampling acceptance for the exponential life distribution had been given in the Military Standards 781C. Algorithms for the numerical determination of the operating characteristics (OC) of the sequential procedures were developed in the 1970's in several papers. In the present lecture, I show analytic methods that can be used to obtain exact values of the OC and the expected sample size (ASN) of two procedures, whose stopping rules are based on prescribed precision requirements.

Plenary Session Speakers

Professor William Q. Meeker
Iowa State University

William Meeker is Professor of Statistics and Distinguished Professor of Liberal Arts and Sciences at Iowa State University. He is a Fellow of the American Statistical Association (ASA), the American Society for Quality (ASQ), and the American Association for the Advancement of Science. He is a past Editor of Technometrics. He is co-author of the books *Statistical Methods for Reliability Data* with Luis Escobar (1998), and the forthcoming second edition of *Statistical Intervals* with Luis Escobar and Gerald Hahn (2017), 14 book chapters, and of numerous publications in the engineering and statistical literature. Bill has won numerous awards for his research as well as the W.G. Hunter Award in 2003 and the ASQ Shewhart medal in 2007. He has done research and consulted extensively on problems in reliability data analysis, warranty analysis, accelerated testing, nondestructive evaluation, and statistical computing.

Title: Statistical Intervals Vive La Differénce!
Chair: **Nalini Ravishanker**, University of Connecticut

Abstract: In 1991, Wiley published *Statistical Intervals: A Guide for Practitioners*, by Gerry Hahn and William Meeker. The main goals of this book were to explain and illustrate the differences among different kinds of statistical intervals (confidence, tolerance, and prediction) and to show how such intervals can be computed under different assumed distributions (e.g., normal, binomial, Poisson, and distribution-free). Due to the success of this book, encouragement from the publisher, and mainly the many technical advances that have been made in the past 25 years, Gerry and Bill, now joined by Luis Escobar, have finished a second edition of this book. In this talk I will briefly review the different kinds of statistical intervals and their uses. The major focus of this talk, however, is to describe some of the many advances made in the past 25 years such as the development of three versatile, general approaches (likelihood-based, bootstrap/simulation-based, and Bayesian-based) to construct statistical intervals for a wide variety of different models, distributions, and types of data. New intervals methods for discrete distributions (binomial and Poisson) and distribution-free intervals will also be reviewed.

Dr. Vijayan N. Nair
Wells Fargo

Vijay Nair is a Statistical Consultant in the Corporate Model Risk Group of Wells Fargo. He was Donald A. Darling Professor of Statistics, Professor of Industrial & Operations Engineering, and Distinguished Scientist in the Michigan Institute for Data Science (MIDAS) at the University of Michigan. He served as Chair of the Statistics Department at Michigan for 12 years. Prior to joining Michigan, he was a Research Scientist at Bell Laboratories in New Jersey for 15 years. He has also consulted extensively with various industries. Vijay has published extensively on statistical methodology, reliability and survival analysis, design of experiments, and engineering statistics. He is a former President of the International Society for Business and Industrial Statistics (ISBIS) and the International Statistical Institute (ISI). He is an elected Fellow of many leading professional societies: American Association for the Advancement of Science, American Statistical Association, American Society for Quality, and the Institute of Mathematical Statistics. He has served as editor of several journals and has chaired numerous committees including the Board of Trustees of the National Institute of Statistical Sciences. He has been at Wells Fargo for the past year.

Title: An Overview of Machine Learning Techniques and Applications in Risk Analysis
Chair: **Haim Bar**, University of Connecticut

Abstract: We will begin with an overview of machine learning (ML) techniques followed by a comparison with computationally intensive statistical methods and the different approaches to modeling. Examples and applications in risk analysis will be used to illustrate the application of ML in banking. These include prediction modeling, variable selection, and anomaly detection.

Sponsors

We are very grateful to the following sponsors of QPRC 2017:

- Minitab
- National Science Foundation, NSF
- SAS/JMP
- Stat-Ease
- TriloByte
- University of Connecticut

Committees

Scientific Program Committee

Martha Gardner, GE

Joseph Glaz, University of Connecticut

Daniel Jeske, UC Riverside

William Q. Meeker, Iowa State University

Nalini Ravishanker, University of Connecticut

Refik Soyer, George Washington University

Emmanuel Yashchin, IBM T. J. Watson Research Center

Organizing Committee

Nalini Ravishanker, University of Connecticut (Chair)

Haim Bar, University of Connecticut (Co-chair)

Xiaojing Wang, University of Connecticut

Jin Xia, GE Global Research

Jian Zou, Worcester Polytechnic Institute

Henry Linder, University of Connecticut (PhD student)

Short Course – Computational Bayesian Methods for Big Data Problems

Steven L. Scott
Director of Statistics Research, Google

Steven Scott is a Director of Statistics Research at Google, where he has worked since 2008. He received his PhD from the Harvard Statistics Department in 1998. He spent 9 years on the faculty of the Marshall School of Business at the University of Southern California. Between USC and Google he also had a brief tenure at Capital One Financial Corp, where he was a Director of Statistical Analysis.

Dr. Scott is a Bayesian statistician specializing in Monte Carlo computation. In his academic life he has written papers on Bayesian methods for hidden Markov models, multinomial logistic regression, item response models, and support vector machines. These methods have been applied to network intrusion detection, web traffic modeling, educational testing, health state monitoring, and brand choice, among others. Since joining Google he has focused on models for time series with many contemporaneous predictors, on scalable Monte Carlo computation, and on Bayesian methods for the multi-armed bandit problem

Monday June 12, 2017, 9:00 – 6:00 pm

This one-day short course will focus on two types of big data problems. The first is regression and time series problems with many predictor variables, the so-called "$p > n$ problem". The course will focus on stochastic search variable selection using spike-and-slab priors, with Google trends data serving as an example where many potential predictors are available. Attendees will learn how to use the 'BoomSpikeSlab' and 'bsts' R packages. Examples that we will consider include model assisted survey sampling, monitoring official statistics, and measuring the impact of market interventions.

The second part of the course focuses on big data problems where the data must be split across multiple machines, where communication between machines is costly. In this case, Consensus Monte Carlo can be used to minimize between-machine communication. Consensus Monte Carlo partitions the data into "shards" assigned to different workers. Each worker runs an independent posterior sampler conditional only on its data shard. The workers then combine their results to form a system-wide "consensus" posterior distribution that approximates the result that would have been obtained if the problem had been handled on a hypothetical single machine. The focus is on the logic of the Consensus Monte Carlo algorithm, for which an R package will be provided. We will not however discuss engineering aspects related to running jobs on multiple machines.

Prerequisites: This course is intended for graduate students, advanced undergraduates, and professionals who have had some prior exposure to Bayesian statistics. Participants should be familiar with Bayes' rule in conjugate normal models, and posterior sampling using Gibbs and MCMC. **Participants should arrive with their own laptop, and with the bsts and BoomSpikeSlab packages installed** (both are available on CRAN).

Short Course Schedule

- 8:30 Registration opens
- 9:00 Short course begins - Session 1
- 10:30 Coffee break
- 11:00 Session 2
- 12:30 PM Lunch and Informal Discussions
- 2:00 PM Session 3
- 3:30 PM Coffee Break
- 4:00 PM Session 4
- 5:30 PM Q&A and Informal Discussions
- 6:00 PM Short Course ends

Short Course participants are invited to an **Informal Pizza Dinner & Chat** from 6:30-8:00 PM. Venue will be announced on June 12.

QPRC 2017 Conference Schedule			
	Tuesday June 13	**Wednesday June 14**	**Thursday June 15**
7:45 AM	Registration Continental Breakfast **OAK**	Continental Breakfast **OAK**	Continental Breakfast **OAK**
8:00 AM			
8:30 AM	Welcome Remarks **OAK 101**	Plenary 2 (Vijayan Nair,Wells Fargo) **OAK 101**	Plenary 3 (William Q. Meeker, Iowa State Univ) **OAK 101**
9:00 AM	Plenary 1 (Shelemyahu Zacks, Binghamton University, Conference Honoree) **OAK 101**		
9:30 AM		AM Break	AM Break
10:00 AM	AM Break	Parallel Invited 4	Parallel Invited 5
10:30 AM	Parallel Invited 1		
11:30 AM			Break
11:45 AM			Parallel Invited 6
12:00 PM	Lunch-Buffet	Lunch-Buffet	
1:00 PM	Software demonstration SAS/JMP **OAK 101**	Software demonstration Stat-Ease **OAK 101**	
1:15 PM			Closing Remarks Boxed Lunch
1:30 PM	Parallel Invited 2	Contributed Parallel	
3:00 PM	PM Break	PM Break	
3:30 PM	Software demonstration Minitab **OAK 101**	Technical Tours (choose one from – BIRC, CHASE, or IMS, all on campus)	
4:00 PM	Parallel Invited 3		
5:00 PM			
6:00 PM		Reception 6:15 Slide show on UConn Tech Park	
6:30 PM	Networking/Cash Bar		
7:00 PM	Banquet (speaker: Nandini Kannan, NSF) **Rome Ballroom**	Reception Poster Session **Student Union Ballroom**	
8:30 PM			

Detailed Conference Program

Tuesday, 10:30AM–12:00PM

Statistical Analysis with Physical Models (OAK 111)

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

1. **Benjamin Haaland**, Georgia Tech.
   "Predicting Solar Irradiance as a Function of Location and Time: Multiple Model Calibration, Non-Stationarity, and Non-Space-Filling Design"
2. **Youngdeok Hwang**, IBM T. J. Watson Research Center
   "Bayesian Pollution Source Identification via an Inverse Physics Model"
3. **Chengrui Li**, Rutgers University-New Brunswick
   "A Sequential Split-Conquer-Combine Approach for Gaussian Process Modeling in Computer Experiments"

Design for Variability at Pratt & Whitney (OAK 112)

Organizer: **Jaime O'Connell**, Pratt & Whitney and **William Q. Meeker**, Iowa State University
Chair: **Jaime O'Connell**, Pratt & Whitney

1. **Jaime O'Connell**, Pratt & Whitney
   "Design for Variation: a Framework for Modern Engineering Design in an Uncertain World"
2. **Benjamin Hall**, Pratt and Whitney
   "An Application in Airfoil High Cycle Fatigue"
3. **Jaime O'Connell**, Pratt & Whitney
   "Leveraging *Design for Variation* to Improve Both Testing and Design A Case Study on Probabilistic Design of Bearings"

Multiple Changepoint Detection in Time Series Data (OAK 117)

Organizer: **Robert Lund**, Clemson University & NSF
Chair: **Nitis Mukhopadhyay**, University of Connecticut

1. **Robert Lund**, Clemson University & NSF
   "Multiple Changepoint Detection in Time Series"
2. **Jonathan Woody**, Mississippi State University
   "A Statistical Analysis of Snow Depth Trends in North America"
3. **Yingbo Li**, Southern Methodist University
   "Bayesian Minimal Description Lengths for Multiple Changepoint Detection"

Tuesday 1:30PM–3:00PM

Applications of Time Series and Spatial Statistics (OAK 111)

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

1. **Gavino Puggioni**, University of Rhode Island
   "Bayesian Mixed Frequency Models for US Unemployment Data Analysis and Forecasting"
2. **Joshua Warren**, Yale University
   "Spatiotemporal Boundary Detection for Localized Smoothing in Areal Data"
3. **Yaohua Zhang**, University of Connecticut
   "Modeling Financial Durations using Estimating Functions"

Recent Advances in Industrial Statistics (OAK 112)

Organizer: **Refik Soyer**, George Washington University
Chair: **Vladimir Pozdnyakov**, University of Connecticut

1. **Ehsan S. Soofi**, University of Wisconsin-Milwaukee
   "Information and Reliability of Escort Models"
2. **Dipak K. Dey**, University of Connecticut
   "Modeling of Large Insurance Claims and Occurrence Data:"
3. **Babak Zafari**, Babson College
   "Modeling First Bid in Retail Secondary Market Online Auctions"

Advances in Statistical Process Control (OAK 117)

Organizer and Chair: **Emmanuel Yashchin**, IBM T. J. Watson Research Center

1. **Wolfgang Schmid**, European University Frankfurt (Oder)
   "Comparison of Joint Control Schemes for Multivariate Normal i.i.d. Output"
2. **Marco Grasso**, Politecnico di Milano
   "Statistical Process Monitoring of Additive Manufacturing via In-situ Sensing"
3. **Axel Gandy**, Imperial College London
   "Calibrating Control Charts when the in Control State is Estimated"

Tuesday 3:30PM–5:30PM

Big Data Applications (OAK 111)

Organizer: **Haim Bar**, University of Connecticut
Chair: **Erin Conlon**, University of Massachusetts-Amherst

1. **Steven L. Scott**, Google
   "Comparing Consensus Monte Carlo Strategies for Distributed Bayesian Computation"
2. **Pierre Lebrun**, Arlenda SA/PharmaLex
   "Statistics, Big Data ... and Small Data"
3. **Haim Bar**, University of Connecticut
   "A Scalable Empirical Bayes Approach to Variable Selection in Generalized Linear Models"
4. **Patrick Flaherty**, University of Massachusetts-Amherst
   "A Deterministic Global Optimization Method for Variational Inference"

Sequential Methods in Quality Control (OAK 112)

Organizer: **Michael Baron**, American University
Chair: **Aleksey Polunchenko**, Binghamton University

1. **Sudeep R. Bapat**, University of Connecticut
   "Purely Sequential Estimation of a Negative Binomial Mean with Applications in Ecology"
2. **Shyamal K. De**, NISER, India
   "Fixed Accuracy Interval Estimation of the Common Variance of Equi-Correlated Normal Distributions"
3. **Yaakov Malinovsky**, University of Maryland, Baltimore County
   "Sequential Estimation in the Group Testing"
4. **Nitis Mukhopadhyay**, University of Connecticut
   "Binomial Non-Sequential and Sequential Sampling Methodologies with Real Applications"

Statistical Methods for Medical Applications (OAK 117)

Organizer: **Dan Jeske**, UC Riverside
Chair: **Maureen Kole**, Applied Research Co.

1. **Tahir Ekin**, Texas State University
   "Improving Medical Audit Quality with Unsupervised Data Mining"
2. **Karel Kupka**, TriloByte
   "PPG - Blood Dynamics Measurement, Modeling and Classification"
3. **Yingtao Bi**, Northwestern University
   "Isoform-Level Quantification from RNA-Seq Data"
4. **Zhanpan Zhang**, GE Global Research
   "Improving Diagnosis of Alzheimers Disease by Data Fusion"

Wednesday 10:00AM–12:00PM

Statistical Applications in Quality for Engineering, Survey and Public Health Data (OAK 104)

Organizers: **Balgobin Nandram and Jian Zou**, Worcester Polytechnic Institute
Chair: **Jian Zou**, Worcester Polytechnic Institute

1. **Jian Zou**, Worcester Polytechnic Institute
   "Dynamic Space-Time Model for Syndromic Surveillance with Particle Filters and Dirichlet Process"
2. **Balgobin Nandram**, Worcester Polytechnic Institute
   "Bayesian Analysis of a Sensitive Small-Area Proportion"
3. **Buddika Peiris**, Worcester Polytechnic Institute
   "Bayesian Analysis of an ROC Curve for Categorical Data Using a Skew-binormal Model"
4. **Adam Ding**, Northeastern University
   "Statistical Modeling for Side-Channel Analysis"

Frontiers in Applied Change-Point Detection (OAK 111)

Organizer: **Aleksey Polunchenko**, Binghamton University
Chair: **Michael Baron**, American University

1. **Marlo Brown**, Niagara University
   "'Detecting Changes in a Poisson Process Monitored at Random Time Intervals"
2. **Aleksey Polunchenko**, Binghamton University
   "Optimal Design of the Shiryaev-Roberts Chart: Give your Shiryaev-Roberts a Headstart"
3. **Vasanthan Raghavan**, Qualcomm
   "Change Propagation Across Sensors: Models, Procedures and Applications"
4. **Grigory Sokolov**, Binghamton University
   "Detecting Change in Correlated Sensor Networks"

Natrella Scholarship Session (OAK 112)

Organizer and Chair: **William Guthrie**, NIST

1. **Yan Wang**, University of Califormia, Los Angeles
   "Statistical Methodology for Data Harmonization"
2. **Simon Mak**, Georgia Institute of Technology
   "Minimax and Minimax Projection Designs using Clustering"

High-Dimensional Data Analysis and Image Testing (OAK 112)

Organizer: **Nalini Ravishanker**, University of Connecticut
Chair: **William Guthrie**, NIST

1. **Ansgar Steland**, RWTH Aachen University, Germany
   "Image Testing in the Presence of Spatial Correlations"

2. **Taras Lazariv**, European University Viadrina, Frankfurt Oder, Germany
"Monitoring of a High-Dimensional Mean Vector"

Wednesday, 1:30PM–3:00PM

Contributed Session 1 (OAK 111)

Chair: **Robert Lund**, Clemson University & NSF

1. **Stephen Clarke**, SABIC Innovative Plastics
"Historical Data Analysis with Autocorrelation"
2. **Caleb King**, Sandia National Laboratories
"Comparison of Accelerated Life Test Plans Based on Exact Small-Sample Methodology and Asymptotic Large-Sample Methodology"
3. **Raymond R. Hill**, Air Force Institute of Technolgoy
"Examining Potential Reductions in Wind Tunnel Testing Data Requirements"
4. **Iulian Ilies**, Northeastern University
"Multimodality in Healthcare Length of Stay and Readmission Time Distributions: Implications on Outlier Detection and Prediction"

Contributed Session 2 (OAK 112)

Chair: **Jeffrey Hooper**, ACC Consumer Finance

1. **Wayne B. Nelson**, Wayne Nelson Statistical Consulting
"An Improved Stockroom Reorder System"
2. **Martin Bezener**, Stat-Ease
"Open Source Tools for Design of Experiments "
3. **Angelo Sant Anna**, Federal University of Bahia
"Monitoring the Low Nonconforming Proportion of High-Quality Process Based on Beta Distribution"
4. **Evgenii Sovetkin**, RWTH University, Aachen, Germany
"Electroluminescence Image Analysis and Suspicious Areas Detection"

Contributed Session 3 (OAK 117)

Chair: **Vincent Raja Anthonisamy**, University of Guyana

1. **Kim Vukovinsky**, Pfizer
"Statistical Contributions to the Pharmaceutical Quality by Design Effort"
2. **Ke Wang**, Pfizer
"Chemical Process Development Experimental Case Studies with Innovative Statistical Tools"
3. **Brent Harrington**, Pfizer
"Analytical Method Development Translating the Analytical Target Profile to the Sampling Strategy"
4. **Abdel-Salam Gomaa**, Qatar University
"Cardiac Surgery Performance Monitoring via Control Charts"

Thursday 10:00AM–11:30AM

Advances in Statistical Methods for Reliability (OAK 104)

Organizer and Chair: **William Q. Meeker**, Iowa State University

1. **Yili Hong**, Virginia Tech.
   "Sequential Test Planning for Polymer Composites"
2. **Maj Jason Freels**, Air Force Institute of Technology
   "R Package SMRD: Comprehensive Life-Data Analysis in R"
3. **Necip Doganaksoy**, Siena College
   "Errors-in-Variables Model in Semiconductor Device Performance and Reliability Assessment"

Design and Analysis for Scientific Forensics and Inverse Problems (OAK 111)

Organizer and Chair: **Christine Anderson-Cook**, Los Alamos National Lab

1. **Lu Lu**, University of South Florida
   "Selecting Top Ranked Solutions for Inverse Prediction with Multiple Responses"
2. **Kevin Quinlan**, Pennsylvania State University
   "Bayesian Design of Experiments for Logistic Regression to Accommodate Multiple Forensic Algorithms"
3. **Christine Anderson-Cook**, Los Alamos National Laboratory
   "Selecting a Discriminating Multivariate Response for Forensic Prediction"

Multivariate Statistical Process Control for Decentralized Wastewater Treatment Facilities (OAK 112)

Organizer and Chair: **Amanda S. Hering**, Baylor University

1. **Tzahi Cath**, Colorado School of Mines
   "Smart Water Reclamation Systems for Tailored Water Reuse and Low Carbon Footprint"
2. **Amanda S. Hering**, Baylor University
   "Comparison of Linear and Nonlinear Dimension Reduction Techniques for Automated Process Monitoring of a Decentralized Wastewater Treatment Facility"
3. **Gabriel J. Odom**, Baylor University
   "Multi-State Multivariate Statistical Process Control"

Thursday 11:45AM–1:15PM

Applications in Business and Industry: An ISBIS Session (OAK 104)

Organizer and Chair: **David Banks**, Duke University

1. **Julie Novak**, IBM T. J. Watson Research Center
   "Revenue Assessment in Large-Scale Businesses"
2. **Beatriz Etchegaray Garcia**, IBM T. J. Watson Research Center
   "Predicting Staffing Needs for Optimal Opportunity Pipeline Management"
3. **David Banks**, Duke University
   "Agent Based Models in Business"

Advances in Stochastic Modeling with Diverse Applications (OAK 111)

Organizer: **Nalini Ravishanker**, University of Connecticut
Chair: **Abdel-Salam Gomaa**, Qatar University

1. **Volodymyr Serhiyenko**, Metabiota
   "Stochastic Modeling of Infectious Diseases"
2. **Vincent Raja Antonisamy**, University of Guyana
   "Reliability Modeling Incorporating Load Share and Frailty"
3. **Vladimir Pozdnyakov**, University of Connecticut
   "Discretely Observed Brownian Motion Governed by a Telegraph Process: Estimation"

Advances in Big Data Methods (OAK 112)

Organizer: **Haim Bar**, University of Connecticut
Chair: **Nandini Kannan**, NSF

1. **Sanguthevar Rajasekaran**, University of Connecticut
   "Big Data Analytics: Basic Algorithms"
2. **Erin Conlon**, University of Massachusetts-Amherst
   "Parallel Markov Chain Monte Carlo Methods for Bayesian Analysis of Big Data"
3. **Kamiar Rahnama Rad**, CUNY
   "Scalable and Robust Model Estimation and Assessment"

Abstracts of Invited Papers

Tuesday, 10:30AM–12:00PM

- **Benjamin Haaland**, Georgia Tech
  "Predicting Solar Irradiance as a Function of Location and Time: Multiple Model Calibration, Non-Stationarity, and Non-Space-Filling Design"
  Benjamin Haaland*, Chih-Li Sung, Wenjia Wang and David Zhao

  We consider the problem of predicting solar irradiance (power per unit area) as a function of location and time using weather station data in addition to data from two weather models. Challenges include very large data size, non-stationarity of the unknown response surface, and non-space-filling weather station locations. Modeling approaches such as local Gaussian process and multi-resolution functional ANOVA, which have potential to work well for this problem, are briefly discussed. We explore a neural network approach to modeling in more depth. Issues include data-driven choice of the number of basis functions, computational efficiency, estimation of the shape, location, and coefficients of basis functions, and inference.

- **Youngdeok Hwang**, IBM T.J. Watson Research Center
  "Bayesian Pollution Source Identification via an Inverse Physics Model"
  Youngdeok Hwang*, Hang Kim and Kyongmin Yeo

  Behavior of air pollution is governed by the complex dynamics, in which air quality of a site is affected by the pollutants transported from neighboring locations via physical processes. To estimate the source of observed pollution, it is crucial to take the atmospheric condition into account. Traditional approach to build empirical models uses observations, but is not able to incorporate the physical knowledge. This drawback becomes particularly severe for the situations where a near-real time source estimation is needed. In this paper, we propose a Bayesian method to estimate the pollution sources, by exploiting both the physical knowledge and observed data. The proposed method uses a flexible approach to utilize the large scale data from the numerical weather prediction model while incorporating the physical knowledge into the model.

- **Chengrui Li**, Rutgers University-New Brunswick
  "A Sequential Split-Conquer-Combine Approach for Gaussian Process Modeling in Computer Experiments"
  Chengrui Li*, Ying Hung and Minge Xie

  Gaussian process (GP) models are widely used in the analysis of computer experiments. However, two critical issues remain unresolved. One is the computational issue in GP estimation and prediction where intensive manipulations of an $n \times n$ correlation matrix are required and become infeasible for large sample size n. The other is how to improve the naive plug-in predictive distribution which is known to underestimate the uncertainty. In this article, we introduce an unified framework that can tackle both issues simultaneously. It consists of a sequential split-conquer procedure, an information combining technique using confidence distributions (CD), and a CD-based predictive distribution. This framework provides estimators and predictors that maintain the same asymptotic efficiency as the conventional method but reduce the computation dramatically. The CD-based predictive distribution contains comprehensive information for statistical inference and provides a better quantification of predictive uncertainty comparing with the plug-in approach. Simulations are conducted to evaluate the accuracy and computational gains. The proposed framework is demonstrated by a data center example based on tens of thousands of computer experiments generated from a computational fluid dynamic simulator.

- **Jaime O'Connell**, Pratt & Whitney
  "Design for Variation: a Framework for Modern Engineering Design in an Uncertain World"
  Jaime O'Connell

  This set of talks describes the purpose and practice of the "Design for Variation" (DFV) discipline at Pratt & Whitney: world leader in the design, manufacture and service of aircraft engines and auxiliary

power units. DFV is a statistical engineering initiative, meaning it aims at melding statistical practice into everyday engineering activities in appropriate ways. Its methods and practices were developed to solve the major challenges faced by many engineers today: 1. How to design physical parts and systems that will meet ever-tightening requirements 2. How to address the various sources of uncertainty and variability designs will experience in real world production and use 3. How to complete a design in less than half the time it traditionally has taken without any reduction in quality. As our solution to these challenges, DFV has enabled us to improve both the fidelity and the speed of the design process. This first talk will focus on introducing the theory and framework of Design for Variation. Then it will review the challenges and successes the team developing and spreading DFV through the organization has experienced. After the first talk, the following talks delivered by Pratt & Whitney engineers will delve into specific applications of the DFV framework and the benefits that have been obtained.

- **Benjamin Hall**, Pratt & Whitney
  "An Application in Airfoil High Cycle Fatigue"
  Benjamin Hall

  This talk will describe the application of Pratt & Whitney "Design for Variation" initiative to the problem of characterizing the risk of airfoil high-cycle fatigue fracture in gas turbine engines. Traditionally, industry has relied on the use of margins as a means for establishing that airfoil designs are robust and therefore that the blades should not be expected to fracture as a result of vibration. While mostly effective, margin-based evaluations are also limited as to the types of questions that can be addressed. As part of Pratt & Whitney's continuing efforts to improve the reliability and performance of its engines, probabilistic approaches to quantifying the risk of airfoil high-cycle fatigue fractures are being developed. The approach is based on a clear and explicit recognition of not only the variability of the airfoils, but the inherent uncertainty that is present in all forecasts of future behavior. To this end, a Bayesian approach is followed in which probabilities are used to quantify states of knowledge. We discuss the advantages and challenges in applying this framework based on experience with several different applications.

- **Jaime O'Connell**, Pratt & Whitney
  "Leveraging 'Design for Variation' to Improve Both Testing and Design A Case Study on Probabilistic Design of Bearings"
  Jaime O'Connell

  In this case study we demonstrate an application of Pratt & Whitney's "Design for Variation" discipline applied to the task of a roller bearing design. The ultimate goal, in this application, was to utilize test data from the "real world" to calibrate a computer model used for design and ensure that roller bearing designs obtained from this model were optimized for maximum robustness to major sources of variation in bearing manufacture and operation. The "Design for Variation" process provides engineers with many useful analysis results even before real world data is applied: high fidelity sensitivity analysis, uncertainty analysis (quantifying a baseline risk of failing to meet design intent) and model verification. The combining of real world data and Bayesian statistical methods that Design for Variation employs to calibrate models, however, goes a step further, validating the accuracy of the model's outputs and quantifying any bias between the model and the real world. As a result of this application, the designers were able to identify the sources of the bias and correct the model's physics-based aspects to more accurately model reality. The improved model is now integrated into all successive bearing design activities. The benefits of this method, which only required a small amount of high quality test data, are now available to all present and future roller bearing designs.

- **Robert Lund**, Clemson University & NSF
  "Multiple Changepoint Detection in Time Series"
  Robert Lund*, Hewa Priyadarshani and Yingbo Li

  This talk presents methods to estimate the number of changepoint time(s) and their locations in time-ordered data sequences. A penalized likelihood objective function is developed from minimum description length (MDL) information theory principles. Optimizing the objective function yields estimates of the changepoint number(s) and location time(s). Our MDL penalty depends on where

the changepoint(s) lie, but not solely on the total number of changepoints (like classical AIC and BIC penalties). Specifically, changepoint configurations that occur relatively closely to one and other are penalized more heavily than sparsely arranged changepoints. The techniques allow for autocorrelation in the observations and mean shifts at each changepoint time. A genetic algorithm, which is an intelligent random walk search, is developed to rapidly optimize the penalized likelihood. Applications to climate series are prominent in the talk.

- **Jonathan Woody**, Mississippi State University
  "A Statistical Analysis of Snow Depth Trends in North America"
  Jonathan Woody

  Several attempts to assess snow depth trends across various regions of North America have been attempted. Previous studies estimated trends by applying various statistical approaches to snow depth data, snow fall data, or their climatological proxies such as snow water equivalents. In most of these studies, inhomogeneities (changepoints) have not been taken into account on a region-wide basis. This talk begins with considerations of how changepoints may effect statistical inference in environmental data, with particular consideration applied towards snow depth observations. A detailed statistical methodology for assessing trends in daily snow depths that accounts for changepoints is considered. Changepoint times are estimated by applying a genetic algorithm to a minimum description length penalized likelihood score. A storage model balance equation with periodic features that allows for changepoints is used to extract standard errors of the estimated trends. The methods are demonstrated on a scientifically accepted gridded data set covering parts of United States and Canada. Results indicate that over half of the grid cells are estimated to contain at least one changepoint and that the average daily snow depth is increasing without changepoints and decreasing with changepoints included in the model.

- **Yingbo Li**, Southern Methodist University
  "Bayesian Minimal Description Lengths for Multiple Changepoint Detection"
  Yingbo Li*, Robert Lund and Anuradha Hewaarachchi

  This paper develops a new class of flexible minimum description length (MDL) procedures for multiple changepoint detection. Existing MDL approaches, which are penalized likelihoods, use data description length information principles to construct penalties that depend on both the number of changepoints and the lengths of the series' segments. While MDL methods have yielded promising results in time series changepoint problems, state-of-the-art MDL approaches are not flexible enough to incorporate domain experts' knowledge that some times are more likely to be changepoints. Furthermore, current MDL methods do not readily handle multivariate series where changepoints can occur in some, but not necessarily all component series. The Bayesian MDL method developed in this paper provides a general framework to account for various prior knowledge, which substantially increases changepoint detection powers. Asymptotically, our estimated multiple changepoint configuration is shown to be consistent. Our method is motivated by a climate application, to identify mean shifts in monthly temperature records. In addition to autocorrelation and seasonal means, our method takes into account metadata, which is a record of station relocations and gauge changes, thus permitting study of documented and undocumented changepoint times in tandem. The multivariate extension allows maximum and minimum temperatures to be jointly examined.

Tuesday 1:30PM–3:00PM

- **Gavino Puggioni**, University of Rhode Island
  "Bayesian Mixed Frequency Models for US Unemployment Data Analysis and Forecasting"
  Gavino Puggioni

  In economic and financial time series applications, explanatory variables that are sampled at higher frequencies than the response are often encountered. For instance, a very common leading indicator for the US unemployment rate, available monthly, is the number of initial claims for unemployment insurance, released weekly by the US Department of Labor. When high frequency covariates are averaged, considerable information can be lost. In this work we propose a Bayesian method based

on a DLM formulation to describe the temporal structure, and a model selection/model averaging procedure that allows an efficient use of all information available at finer frequencies. Different choices of prior distributions for model space, including dilution priors, are discussed. By taking into account parameter and model uncertainty, the results show an improvement in quality of fit and forecasting accuracy.

- **Joshua Warren**, Yale University
  "Spatiotemporal Boundary Detection for Localized Smoothing in Areal Data"
  Joshua Warren*, Samuel Berchuck and Jean-Claude Mwanza

  Boundary detection is a technique used to identify discrete changes in otherwise smooth spatial surfaces for spatially referenced areal data. These methods are well-established for areal data occurring at a single point in time. However, spatially referenced datasets often include repeated measures in the form of longitudinal observations, inducing a potentially complex spatiotemporal correlation structure among the data. There is a need to develop new statistical methodology to analyze boundaries in a spatial surface and their changes across time. Motivated by a challenging problem in glaucoma research, we develop a Bayesian hierarchical model that allows for localized spatial structure, while incorporating a flexible modeling framework for repeated measures. The effectiveness of this approach is assessed in simulation and then applied to a longitudinal series of visual fields used for diagnosing glaucoma progression. Comparisons with an existing non-temporal model are also included. We use data from the Vein Pulsation Study Trial in Glaucoma and the Lions Eye Institute trial registry, Perth, Western Australia.

- **Yaohua Zhang**, University of Connecticut
  "Modeling Financial Durations using Estimating Functions"
  Yaohua Zhang*, Jian Zou, Nalini Ravishanker and Aerambamoorthy Thavaneswaran

  Accurate modeling of patterns in inter-event durations is of considerable interest in high-frequency financial data analysis. The class of logarithmic autoregressive conditional duration (Log ACD) models provides a rich framework for analyzing durations, and recent research is focused on developing fast and accurate methods for fitting these models to long time series of durations under least restrictive assumptions. This article describes an optimal semi-parametric modeling approach using martingale estimating functions. This approach only requires assumptions on the first few conditional moments of the durations and does not require specification of the probability distribution of the process. We introduce three approaches for parameter estimation in our methodology, including solution of nonlinear estimating equations, recursive formulas for the vector-valued parameter estimates, and iterated component-wise scalar recursions. All three estimation methods are compared via extensive numerical studies. Effective starting values from an approximating time series model increase the accuracy of the final estimates. We demonstrate our approach via a simulation study and a real data illustration based on high-frequency transaction level data on several stocks.

- **Ehsan S. Soofi**, University of Wisconsin-Milwaukee
  "Information and Reliability of Escort Models"
  Majid Asadi, Nader Ebrahimi and Ehsan S. Soofi*

  In physics literature, the normalized power of one or two probability density or mass functions are called escort and generalized escort distributions. Probability models that are normalized powers of two density functions and powers of one or two survival functions have been used in statistics, reliability, and survival analysis for decades in various contexts and with various names or nameless. We synthesize the exiting formulations that provide the escort of densities and explore the hazard rate properties of the generalized escort models. A notable property is that the generalized escort of densities of two non-constant hazard distributions can be a constant hazard rate model. We also note that the survival function of the proportional hazards model is the escort of the baseline survival function and the survival function of the mixture hazard rates model is the generalized escort of two survival functions. We show that these models are characterized by various information theoretic and variation formulations. Potential applications of the generalized escort include models for lifetimes of items with heterogeneous quality.

- **Dipak K. Dey**, University of Connecticut
  "Modeling of Large Insurance Claims and Occurrence Data"
  Dipak K. Dey

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

- **Babak Zafari**, Babson College
  "Modeling First Bid in Retail Secondary Market Online Auctions"
  Babak Zafari* and Refik Soyer

  The online commerce has greatly changed the trading markets for both businesses and consumers. A big part of this environment is the way online auctions are conducted. In this work, we propose a Bayesian dynamic probit and a beta regression model to describe participation and time to first bid in retail secondary market online auctions. For a given auction, we also predict the bidder who will place the first bid earlier than other participants. In developing the models, we consider some auction-specific and bidder-specific explanatory variables. The bidder-specific characteristics take into account both bidders' bidding behavior and their experience based on their previously participated auctions and their current activities at the time of the bid.

- **Wolfgang Schmid**, European University Frankfurt (Oder)
  "Comparison of Joint Control Schemes for Multivariate Normal i.i.d. Output"
  M. Morais, W. Schmid*, P. Ramos, T. Lazariv and A. Pacheco

  The performance of a product frequently relies on more than one quality characteristic. In such a setting, joint control schemes are used to determine whether or not we are in the presence of unfavorable disruptions in the location and spread of a vector of quality characteristics. A common joint scheme for multivariate output comprises two constituent control charts: one for the mean vector based on a weighted Mahalanobis distance between the vector of sample means and the target mean vector; another one for the covariance matrix depending on the ratio between the determinants of the sample covariance matrix and the target covariance matrix. Since we are well aware that there are plenty of quality control practitioners who are still reluctant to use sophisticated control statistics, this paper tackles Shewhart-type charts for the location and spread based on a few pairs of control statistics that depend on the nominal mean vector and covariance matrix. We recall or derive the joint probability density functions of these pairs of control statistics in order to investigate the impact on the ability of the associated joint schemes to detect shifts in the process mean vector or covariance matrix for various out-of-control scenarios.

- **Marco Grasso**, Politecnico di Milano
  "Statistical Process Monitoring of Additive Manufacturing via in-Situ Sensing"
  Marco Grasso* and Bianca Maria Colosimo

  Thanks to rapid technological advances, metal Additive Manufacturing (AM) systems are becoming more and more suitable to produce functional parts, ranging from one-of-a-kind to series productions. However, the lack of process repeatability and stability still represents a barrier for the industrial breakthrough of these processes. Indeed, the major AM potentialities are particularly attractive in industrial sectors like aerospace and bio-medical, where defects avoidance is of fundamental importance. Several authors and the main metal AM system developers pointed out the need to develop and implement in-situ monitoring tools able to keep under continuous control the stability of the process. Most efforts in the literature and in industry have been focused so far on gathering data during the process via in-situ sensing. However, there still a lack of methods to make sense of big and fast in-situ

data streams and to quickly detect the onset of process defects during the layer-wise production of a part. This study deals with the problem of in-situ monitoring of metal AM processes from a statistical data analysis perspective. First, different case studies in Selective Laser Melting (SLM), a category of metal AM technologies, are presented, where in-situ gathered data consists of image streams captured via machine vision equipment. A suite of statistical process control (SPC) techniques is then presented to design control charting schemes able to automatically signal deviations from in-control patterns. The proposed approaches aim at extracting the relevant information content from image data, by describing the spatial or spatio-temporal patterns in order to determine both when and where a defect has originated. They exploit and couples image processing and segmentation techniques, multivariate data analysis and dimensionality reduction tools to synthesize the information content and characterize the 'signatures' of the SLM process. A discussion of major challenges imposed by metal AM applications and existing open issues is provided to highlight the most relevant directions of future research efforts in this field.

- **Axel Gandy**, Imperial College London
  "Calibrating Control Charts when the in Control State is Estimated"
  Axel Gandy* and Jan Terje Kvaloy

  In practical applications of control charts the in-control state and the corresponding chart parameters are usually estimated based on some past in-control data. The estimation error then needs to be accounted for. We discuss how this can be accounted for in various situations,with a particular focus on bootstrap methods - and how this has been implemented in an R-package.

Tuesday 3:30PM–5:30PM

- **Steven L. Scott**, Google
  "Comparing Consensus Monte Carlo Strategies for Distributed Bayesian Computation"
  Steven L. Scott

  Consensus Monte Carlo is an algorithm for conducting Monte Carlo based Bayesian inference on large data sets distributed across many worker machines in a data center. The algorithm operates by running a separate Monte Carlo algorithm on each worker machine, which only sees a portion of the full data set. The worker-level posterior samples are then combined to form a Monte Carlo approximation to the full posterior distribution based on the complete data set. We compare several methods of carrying out the combination, including a new method based on approximating worker-level simulations using a mixture of multivariate Gaussian distributions. We find that resampling and kernel density based methods break down after 10 or sometimes fewer dimensions, while the new mixture-based approach works well, but the necessary mixture models take too long to fit.

- **Pierre Lebrun**, Arlenda SA / PharmaLex
  "Statistics, Big Data... and Small Data"
  Pierre Lebrun

  Nowadays, big data is more than buzz words, and many people succeed implementing it at very large scale. However, a considerable amount of the tools and methodologies that are used remains the same for decades. GLM Regressions, SVM, random forest, deep and reinforcement learning, etc. Most of these methods perform well given the context they are used, but they suffer for lack of inclusion of uncertainty in their prediction. For instance, they generally miss the point to answer what is the chance that a result is due to poor model fitting? In statistics providing prediction accounting for variability remains one of the hot topic, especially with moderate to complex models. Solutions are known, in frequentist and Bayesian statistics for classical Normal model, but when the problem is more intricate, things become more difficult. However, using the Bayesian framework, or at the very least Bayesian thinking, allows providing results accounting for this predictive uncertainty, even when no classical solution exists. Solutions for big data is nevertheless challenging given the Bayesian simulation framework, that has a computer costs that one cannot neglect. Through some examples, it will be shown how some classical questions related to process control and time series analysis can be answered through exact solution and approximate acceptable solutions.

- **Haim Bar**, University of Connecticut
  "A Scalable Empirical Bayes Approach to Variable Selection in Generalized Linear Models"
  Haim Bar*, James Booth and Martin Wells

  A new empirical Bayes approach to variable selection in the context of generalized linear models is developed. The proposed algorithm scales to situations in which the number of putative explanatory variables is very large, possibly much larger than the number of responses. The coefficients in the linear predictor are modeled as a three-component mixture allowing the explanatory variables to have a random positive effect on the response, a random negative effect, or no effect. A key assumption is that only a small (but unknown) fraction of the candidate variables have a non-zero effect. This assumption, in addition to treating the coefficients as random effects facilitates an approach that is computationally efficient. In particular, the number of parameters that have to be estimated is small, and remains constant regardless of the number of explanatory variables. The model parameters are estimated using a modified form of the EM algorithm which is scalable, and leads to significantly faster convergence compared with simulation-based fully Bayesian methods.

- **Patrick Flaherty**, University of Massachusetts-Amherst
  "A Deterministic Global Optimization Method for Variational Inference"
  Patrick Flaherty

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

- **Sudeep R. Bapat**, University of Connecticut
  "Purely Sequential Estimation of a Negative Binomial Mean with Applications in Ecology"
  Nitis Mukhopadhyay and Sudeep R. Bapat*

  We discuss a set of purely sequential strategies to estimate an unknown negative binomial mean under different forms of loss functions. We develop point estimation techniques where the thatch parameter may be known or unknown. Both asymptotic first-order efficiency and risk efficiency properties will be elaborated. The results will be supported by an extensive set of data analysis carried out via computer simulations for a wide variety of sample sizes. We observe that all of our purely sequential estimation strategies perform remarkably well under different situations. We also illustrate the implementation of these methodologies using real data-sets from ecology, namely, weed count data and data on migrating woodlarks.

- **Shyamal K. De**, NISER, India
  "Fixed Accuracy Interval Estimation of the Common Variance of Equi-Correlated Normal Distributions"
  Shyamal K. De* and Nitis Mukhopadhyay

  Correlated multivariate normal distributions appear in many areas of statistics such as MANOVA and repeated measures ANOVA. For instance, in MANOVA with m-dimensional response variable, one can model the error part by m-dimensional normal distribution with mean zero and some covariance matrix. A popular choice of covariance matrix is a correlation matrix with equal off-diagonal entries multiplied by a common variance parameter. Many researchers investigated the problem of interval estimation of the common variance parameter in the presence of the nuisance common correlation parameter. In this talk, we will propose two-stage sampling methodology to construct fixed-accuracy

confidence intervals for the common variance parameter. We will derive exact distributions of the stopping variable and the estimator of the common variance at stopping. The coverage probabilities of the proposed interval estimator will be computed exactly and will be shown to be nearly the same as the prescribed level. Moreover, we will establish the asymptotic second-order efficiency and asymptotic consistency properties of the proposed methodology.

- **Yaakov Malinovsky**, University of Maryland, Baltimore County
  "Sequential Estimation in the Group Testing"
  Gregory Haber, Yaakov Malinovsky* and Paul Albert

  Estimation using pooled sampling has long been an area of interest in the group testing literature. Such research has focused primarily on the assumed use of fixed sampling plans (i), although some recent papers have suggested alternative sequential designs that sample until a predetermined number of positive tests (ii). One major consideration, including in the new work on sequential plans, is the construction of debiased estimators which either reduce or keep the mean square error from inflating. Whether, however, under the above or other sampling designs unbiased estimation is in fact possible has yet to be established in the literature. In this work, we introduce a design which samples until a fixed number of negatives (iii), and show that an unbiased estimator exists under this model, while unbiased estimation is not possible for either of the preceding designs (i) and (ii). We present new estimators under the different sampling plans that are either unbiased or that have reduced bias relative to those already in use as well as generally improve on the mean square error.

- **Nitis Mukhopadhyay**, University of Connecticut-Storrs
  "Binomial Non-Sequential and Sequential Sampling Methodologies with Real Applications"
  Nitis Mukhopadhyay

  The binomial non-sequential and sequential sampling methods are carried out in practice when the response variable from a sampling unit is dichotomous and the employed design is simple random sampling with replacement (SRSWR). When a population's size is large, simple random sampling without replacement (SRSWOR) is customarily approximated with SRSWR that essentially leads one back to binomial sampling, especially when the sample size is small compared with population size. I will first summarize preliminary ideas involving non-sequential binomial sampling with an emphasis on concepts and examples from confidence intervals. I will then discuss a court-case in "computer fraud" that was launched more than 12 years ago. With the help of figures-charts-analysis, I will exhibit how simple ideas of implementing appropriate binomial sequential sampling methodologies could have saved hundreds of thousands of dollars!

- **Tahir Ekin**, Texas State University
  "Improving Medical Audit Quality with Unsupervised Data Mining"
  Tahir Ekin

  It is estimated that three to ten percent of the annual medical care spending results in overpayments. Medical audits are utilized on a sample of claims to investigate the legitimacy of the submissions and identify overpayments. This talk discusses the use of unsupervised data mining approaches as pre-screening tools to improve the quality of the medical audits. Data mining is used to identify the hidden patterns among providers and medical procedures and to find outliers. It can help the auditors to devote the investigation resources on claims that are more likely to be overpaid. We illustrate the utilization of the proposed methods using U.S. Medicare medical claims data.

- **Karel Kupka**, TriloByte
  "PPG - Blood Dynamics Measurement, Modeling and Classification"
  Karel Kupka*, Jaroslav Jansa and Martin Sikora

  Haemodynamics data carries information about both central and peripheral arterial system. Photoplethysmography (PPG) is a non-invasive method measuring changes in light absorption in peripheral tissue as a function of time. It is widely used as a simple diagnostic tool to measure heart beat rate. In this contribution we investigate univariate filtering and smoothing algorithms and orthogonal linear harmonic regression models to parametrize the PPG pulse waveform shapes. In the parametric space we employ and compare supervised classification methods such as Fischer LDA, Neural network

classifier, Mahalanobis classifier, Support Vector Machine Classifier to aid patient diagnosis based on different PPG curve shapes. Unsupervised clustering was also used to suggest possible groups of risk factors. We also investigate alternative method called baroplethysmography (BPG) which records relative blood pressure signal simultaneously with the PPG. Both measurements are taken from the patient's fingers. We conclude that both measurement methods may provide information represented by independent parameters in multidimensional (typically 15-20 dimensions) parametric space. which can be used both as long-term stability monitoring of a particular patient and to classify different patients with respect to their arterial system state and related diseases, possibly including early warning tool for arteriosclerosis or diabetes.

- **Yingtao Bi**, Northwestern University
  "Isoform-Level Quantification from Rna-Seq Data"
  Yingtao Bi* and Ramana V. Davuluri

  Recent genome-wide studies have discovered that majority of human genes produce multiple transcript-variants/protein-isoforms. Several diseases, including cancers, have been associated with dysregulation of alternative splicing. RNA-seq technology has been extensively used to infer isoform abundance. In this talk we present an evaluation study of various RNA-seq data analysis tools for isoform level expression quantification. We also discuss a novel proposed statistical method to identify differential isoform usage from RNA-seq data.

- **Zhanpan Zhang**, GE Global Research
  "Improving Diagnosis of Alzheimer's Disease by Data Fusion"
  Zhanpan Zhang, Dipen P. Sangurdekar, Susan Baker and Cristina A. Tan Hehir

  Blood-based protein biomarkers predicting brain amyloid burden would have great utility for the enrichment of Alzheimer's Disease (AD) clinical trials, including large-scale prevention trials. We adopt data fusion to combine multiple high dimensional data sets upon which classification models are developed to predict amyloid burden as well as the clinical diagnosis. Specifically, non-parametric techniques are used to pre-select variables, and random forest and multinomial logistic regression techniques with LASSO penalty are performed to build classification models. We apply the proposed data fusion framework to the AIBL imaging cohort and demonstrate improvement of the clinical status classification accuracy. Furthermore, variable importance is evaluated to discover potential novel biomarkers associated with AD.

Wednesday 10:00AM–12:00PM

- **Jian Zou**, Worcester Polytechnic Institute
  "Dynamic Space-Time Model for Syndromic Surveillance with Particle Filters and Dirichlet Process"
  Hong Yan, Zhongqiang Zhang and Jian Zou*

  Massive spatio-temporal data are challenging for statistical analysis due to their low signal-to-noise ratios and high-dimensional spatio-temporal structure. To resolve these issues, we propose a novel Dirichlet process particle filter (DPPF) model. The Dirichlet process models a set of stochastic functions as probability distributions for dimension reduction, and the particle filter is used to solve the nonlinear filtering problem with sequential Monte Carlo steps where the data has a low signal-to-noise ratio. Our data set is derived from surveillance data on emergency visits for influenza-like and respiratory illness (from 2008 to 2010) from the Indiana Public Health Emergency Surveillance System. Numerical results show that our model significantly improves the outbreak detection performance in real data analysis.

- **Balgobin Nandram**, Worcester Polytechnic Institute
  "Bayesian Analysis of a Sensitive Small-Area Proportion"
  Balgobin Nandram

  College cheating is a serious problem in the US, and one way to get information about it is to use indirect questioning. In sample surveys with sensitive items, sampled individuals either do not respond (nonignorable nonresponse) or they respond untruthfully. For example, respondents usually give

negative answers to sensitive items when the responses should actually have been positive, thereby leading to an estimate of the sensitive population proportion that is too small. For this study, we have sparse binary data on college cheating, collected under an unrelated-question design, from several locations (small areas) on a US campus. We have used a hierarchical Bayesian model to capture the variation in the observed binomial counts from these locations and to estimate the sensitive proportions for all locations. For our application on college cheating, there are significant reductions in the posterior standard deviations of the sensitive proportions under the small-area model in comparison to an analogous individual-area model. A simulation study confirms the gain in precision and, surprisingly, shows the estimates under the small-area model are closer to the truth than the corresponding estimates under the individual-area model.

- **Buddika Peiris**, Worcester Polytechnic Institute
  "Bayesian Analysis of an ROC Curve for Categorical Data Using a Skew-binormal Model"
  Buddika Peiris

  In a taste-testing experiment, foods are withdrawn from storage at various times and a panel of tasters are asked to rate the foods, which are rated on a nine-point hedonic scale. We provide a statistical procedure which can assess the difference between fresh foods and foods withdrawn a few months later. Thus, we have two sets of ordinal data, one for the fresh foods and the other for the foods which are withdrawn. A natural and popular way to compare two withdrawals is to use the receiver operating characteristic (ROC) curve and the area under the curve (AUC). We perform Bayesian methods, which incorporate a stochastic ordering, to obtain the AUC, but these methods are well known. However, our first method, which robustifies the binormal model, is novel. our second method is more innovative because it uses a skew binormal model with additional robustification like the binormal method. We use the Gibbs sampler to fit both models in order to estimate the ROC curves and the AUCs. These AUCs demonstrate that there is not much difference between fresh foods and those withdrawn later using both methods. However we have shown, using marginal likelihoods, that skew binormal model is better.

- **Adam Ding**, Northeastern University
  "Statistical Modeling for Side-Channel Analysis"
  Adam Ding

  Side-channel attacks (SCA) are an emerging threat to embedded systems. The attackers use physical side-channel measurements, such as power consumption or the execution time, to break theoretical secure cryptographic algorithms in the implemented systems. We discuss the applications of statistical models and statistical methods in evaluation and detection of side-channel leakage. We develop a statistical model that explicitly express effects of the physical system noise and the crypto algorithmic property on the SCA security. Metrics are developed to evaluate SCA resistance of the physical systems. Multivariate leakage in masking protected systems are also modeled and studied with statistical methods.

- **Marlo Brown**, Niagara University
  "Detecting Changes in a Poisson Process Monitored at Random Time Intervals"
  Marlo Brown

  We look at a Poisson process where the arrival rate changes at some unknown time point. We monitor this process only at certain time points. At each time point, we count the number of arrivals that happened in that time interval. In previous work, it was assumed that the time intervals were fixed in advanced. We relax this assumption to assume that the time intervals that the process is monitored is also random. For a loss function consisting of the cost of late detection and a penalty for early stopping, we develop, using dynamic programming, the one and two steps look ahead Bayesian stopping rules. We then compare various observation schemes to determine the best model. We provide some numerical results to illustrate the effectiveness of the detection procedures.

- **Aleksey Polunchenko**, Binghamton University
  "Optimal Design of the Shiryaev-Roberts Chart: Give your Shiryaev-Roberts a Headstart"
  Aleksey Polunchenko

It is well-known from the seminal 1982 paper by J.M. Lucas and R.B. Crosier published in Technometrics that giving a CUmulative SUM (CUSUM) inspection scheme a headstart provides an additional degree of freedom (in addition to the control limit) wherewith the scheme's performance can not only be improved but also be customized more flexibly than through the control limit alone. This work is an attempt to extend the results obtained by J.M. Lucas and R.B. Crosier for the CUSUM scheme to the latter's far less famous yet rather capable competitor – the Shiryaev-Roberts (SR) chart. More specifically, we seek to understand the effect of headstarting on the performance of the SR chart. Through an extensive numerical optimization of the SR chart carried out for the standard Gaussian mean-shift scenario, we demonstrate, just as J.M. Lucas and R.B. Crosier did for the CUSUM scheme, that giving the SR chart a headstart, too, boosts up the chart's performance, and, in particular, equips it with a Fast Initial Response (FIR) feature. However, and this is the main result of this work, starting the SR chart off a carefully designed initial value turns out to be akin to "putting the SR chart on steroids": the SR chart becomes not just faster, it becomes almost the fastest, i.e., almost the best one can do when the process under surveillance is equally likely to go out of control at any time moment. Needless to say, the CUSUM scheme is not as efficient. To make our findings practically useful, we tabulate the obtained optimal headstart, control limit, and the corresponding "worst possible" out-of-control Average Run Length (ARL), considering mean shifts of diverse magnitudes (from faint to contrast) and a wide range of levels of the in-control ARL.

- **Vasanthan Raghavan**, Qualcomm
  "Change Propagation Across Sensors: Models, Procedures and Applications"
  Vasanthan Raghavan

  Low-cost sensors are often used to monitor certain signature changes in diverse applications such as civil engineering (bridge monitoring), healthcare (concussion monitoring), etc. While the assumption that all the sensors observe the same change process at the same time results in a ready extension of the single-sensor change detection approaches to this setting, this assumption is naive as it does not capture real-world signatures accurately. A recent take on handling this deficiency is to assume that an unknown subset of sensors observe the same/different set of change processes, but all at the same time. An alternate take on this problem is to assume that all the sensors observe change "eventually", but the change propagates across the sensors as a Markovian process. The focus of this talk is on a comparative study of the two modeling approaches and the ramifications on detecting changes with either model.

- **Grigory Sokolov**, SUNY at Binghamton
  "Detecting Change in Correlated Sensor Networks"
  Grigory Sokolov* and Georgios Fellouris

  A random process undergoes a statistical change at some unknown time. The pre-change behavior is assumed to be known, but after the change its distribution is only known to be one of finitely many. We consider the problem of detecting the time of change as soon as possible, while controlling the rate of false alarms. In this talk we will review rules whose additional detection delay—relative to the one that could be achieved if the post-change distribution was known—remains bounded as the rate of false alarms goes to zero. Finally, we will present a correlated sensor-network example that fits into the framework, and consider the case when certain constraints are imposed on inter-sensor communications.

- **Yan Wang**, University of California at Los Angeles
  "Statistical Methodology for Data Harmonization"
  Yan Wang*, Honghu Liu

  In the age of big data, multiple data sources are intended to be merged together in order to make inferences and answer complicated questions. If done appropriately, the process can foster novel opportunities for research with richer database and more powerful results. Data harmonization refers to the adjustment of differences and inconsistencies of data sets among the different measurements, methods, procedures, schedules, specifications, or systems to make the data sources uniform or compatible. It supports pooling the comparable data source to increase accuracy and efficiency. It is the essential

step to turn big data into thick information. We review the statistical methods used in literature for data harmonization, e.g. standardized score, re-categorization, missing value imputation, latent variable method. We propose a two-stage method for data harmonization process. The first stage is at measurement level to create a common measure based on statistical theory, e.g. item response theory (IRT) to create the latent common measure. The second stage is at the study level to take the design features among different data resources into consideration. A harmonization circle is invented to visualize the harmonization process. A harmonization score based on metric theory is created to quantify the harmonization process and to make inferences on data harmonization. The harmonization method and metric theory is applied to a multisite collaboration of HIV adherence study among sixteen institutions to create the common measure and evaluate the harmonization process. In the study, adherence is measured by different recall intervals of self-reported adherence and electronic drug monitoring (EDM).

- **Simon Tsz Fung Mak**, Georgia Institute of Technology
  "Minimax and Minimax Projection Designs using Clustering"
  Simon Tsz Fung Mak*, V. Roshan Joseph

  Minimax designs provide a uniform coverage of a design space by minimizing the maximum distance from any point in this space to its nearest design point. While minimax designs have many useful applications, e.g., for optimal sensor allocation or as space-filling designs for computer experiments, there has been little work in developing algorithms for generating these designs, due to its computational complexity. In this talk, we propose a new hybrid algorithm combining particle swarm optimization and clustering for efficiently generating minimax designs on any convex and bounded design space. Simulation studies show that the proposed algorithm provides improved minimax performance over existing methods on a variety of design spaces. Finally, we introduce a new type of experimental design called a minimax projection design, and show that this proposed design provides better minimax performance on projected subspaces compared to existing designs. The usefulness of minimax and minimax projection designs is then illustrated using a real-world example on rocket injector design.

- **Ansgar Steland**, RWTH Aachen University, Germany
  "Image Testing in the Presence of Spatial Correlations"
  Ansgar Steland

  Image data is of increasing importance in quality control, e.g. to control and monitor the quality of materials such as silicon wafers used to manufacture CPUs or solar cells. We discuss several approaches and new results, in order to statistically evaluate whether image data complies with given specifications. As digitized images given by random matrices may be affected by spatial correlations, the estimation of the variance of the sum (or mean) of an image is crucial for proper standardization, but challenging in practice. We discuss new results on its estimation by a threshold-type approach. Simulations show that the null distributions of the corresponding t-type test statistics can be very well approximated when using that new estimator.

- **Taras Lazariv**, European University Viadrina, Frankfurt Oder, Germany
  "Monitoring of a High-Dimensional Mean Vector"
  Taras Lazariv*, Nestor Parolya, Wolfgang Schmid

  The aim of this paper is to develop new control charts for monitoring a high-dimensional mean vector using large dimensional random matrix theory. We construct control charts for the case when both the dimension p and the sample size n of the data generating process tend to infinity under the concentration condition $p/n \to c \in (0, \infty]$. The results are derived using recently published results on the inference of high-dimensional mean vectors. Via an extensive simulation study we note a superior performance of the new control chart in comparison to the traditional and state-of-the-art benchmarks in the case of sparse covariance matrices.

Thursday 10:00AM–11:30AM

- **Yili Hong**, Virginia Tech
  "Sequential Test Planning for Polymer Composites"
  I-Chen Lee, Yili Hong*, ST Tseng and Tirthankar Dasgupta

  Polymer composite materials are widely used in areas such as aerospace and alternative energy industries, due to their lightweight and comparable levels of strength and endurance. To ensure the material can last long enough in the field, accelerated cyclic fatigue tests are commonly used to collect data and then make prediction for the field performance. While there has been a lot development in optimum test planning, most of the methods assume that the true parameter values are known. However, in reality, the true model parameters may depart from the planning values. In this paper, we propose a sequential test planning strategy for polymer composites. We use Bayesian framework for the sequential model updating. We also use extensive simulation to evaluate the properties of the proposed sequential test planning strategy.

- **Maj Jason Freels**, Air Force Institute of Technology
  "R Package SMRD: Comprehensive Life-Data Analysis in R"
  William Meeker, Luis Escobar, Jason Freels*

  Since its publication in 1998, Statistical Methods for Reliability Data by W. Q. Meeker and L. A. Escobar has been recognized as a foundational resource in analyzing time to event data. Along with the text, the authors provided an S-Plus software package, called SPLIDA, to help readers utilize the methods presented in the text. Today, R is the most popular statistical computing language in the world, largely supplanting S-Plus. The SMRD package is the result of a multi-year effort to completely rebuild SPLIDA, to take advantage of the improved graphics and workflow capabilities available in R. This presentation introduces the SMRD package, outlines the improvements and shows how the package works seamlessly with the rmarkdown and shiny packages to dramatically speed up your workflow. The presentation concludes with a discussion on what improvements still need to be made prior to publishing the package on the CRAN

- **Necip Doganaksoy**, Siena College
  "Errors-in-Variables Model in Semiconductor Device Performance and Reliability Assessment"
  Necip Doganaksoy

  Designers and manufacturers of battery powered mobile electronics are concerned with the increased propensity for current leakage as devices get smaller in size. As a result, assessment of device performance involves joint analysis of measurement data on saturation drain current (a measure of performance) and off-state current (leakage, a measure of reliability). In this paper, we extend the traditional errors-in-variables model to compare device reliability and performance under different design and processing conditions.

- **Lu Lu**, University of South Florida
  "Selecting Top Ranked Solutions for Inverse Prediction with Multiple Responses"
  Lu Lu*, Christine M. Anderson-Cook

  Inverse prediction utilizes relationships between multiple inputs and outputs from a process to predict the most likely set of inputs (causes) that may have generated a set of newly observed outputs. It has important applications in contraband forensics and reverse detection problems. Current practice focuses on seeking a single best match for the inputs based on the observed output responses by minimizing a measured distance between the new observation and the estimated responses at different locations. This talk presents an alternative approach, which uses Pareto Front optimization to identify multiple leading solutions ranked with some quantification of robustness to different prioritizations of responses. The impact of parameter estimation uncertainty on inverse prediction is also considered when choosing top ranked solutions. We present results from a simulation study to demonstrate consistently higher success rates for identifying true input locations that generated the new observations as well as reduced discrepancy between the true and identified locations.

- **Kevin Quinlan**, Pennsylvania State University-Main Campus
  "Bayesian Design of Experiments for Logistic Regression to Accommodate Multiple Forensic Algorithms"
  Kevin Quinlan*, Christine Anderson-Cook, Kary Myers

  When evaluating the performance of several Forensic classification algorithms, it is desirable to create a design that considers a variety of performance levels across the algorithms. We describe a strategy to use Bayesian design of experiments with multiple prior estimates to capture anticipated performance. Our goal is to characterize results from the different algorithms as a function of different explanatory variables and use this to help choose a design about which units to test. Bayesian design of experiments has been shown to be very successful for generalized linear models, including logistic regression models. We develop methodology for the case where there are several potentially non-overlapping priors under consideration. The Weighted Priors method performs well for a broad range of true underlying model parameter choices and is more robust when compared to other reasonable design choices. Additionally we show how this can be applied in the multivariate input case and provide some useful summary measures. We illustrate the method with several examples.

- **Christine Anderson-Cook**, Los Alamos National Laboratory
  "Selecting a Discriminating Multivariate Response for Forensic Prediction"
  Christine Anderson-Cook*, Edward V. Thomas, John R. Lewis, Tom Burr, Michael S. Hamada

  Inverse prediction is important in a wide variety of scientific and engineering contexts. One might use inverse prediction to predict fundamental properties/characteristics of an object/material by using multiple measurements obtained from it. This can be accomplished by inverting parameterized forward models that relate the measurements (responses) to the properties/characteristics of interest. Sometimes forward models are science based; but often, forward models are empirically based, using the results of experimentation. While nature dictates the causal relationship between factors and responses, experimenters can influence control of the complexity, accuracy, and precision of forward models that can be constructed through selection of factors, factor levels, and the designed experiment. Simple analysis and examples are used to illustrate how one might select an informative and discriminating subset of response variables among candidates in cases where the number of response variables available is limited by difficulty, expense, and/or availability of material.

- **Tzahi Cath**, Colorado School of Mines
  "Smart Water Reclamation Systems for Tailored Water Reuse and Low Carbon Footprint"
  Tzahi Cath

  Life and society rely on adequate water supplies to sustain humans and the environment. Water planners and engineers must look beyond traditional methods of water supply (e.g., infrastructure developments and inter-basin water transfers) and adopt an integrated, whole system approach to managing water assets. These assets must include locally available reclaimed water as a strategic supply for balancing urban water use, meeting short-term needs, and improving long-term supply reliability. The integration of onsite, decentralized wastewater treatment systems into existing urban water infrastructure is an attractive option for recovering water, nutrients, and other resources locally for multi-purpose reuse. This is true for both industrial and municipal applications. To facilitate wastewater treatment and encourage reuse tailored to local needs, smart hybrid treatment processes and systems are needed. These hybrid treatment systems must be designed for flexibility in order to meet product water (effluent) quantity and quality that can be tailored for seasonal or site specific uses (i.e., nutrients may be retained in the effluent depending on need) rather than just meeting discharge standards. However, operating unattended small, distributed wastewater treatment systems requires quick response to process failure in order to avoid discharge of polluted water to the environment. Furthermore, biological processes are more prone to failure, and restoring their performance after failure is lengthy and complex. Therefore, simple early detection systems to alert operators to potential process failure can prevent environmental damage and avoid costly restoration. Using existing process monitoring equipment and data acquisition systems in conjunction with statistical models can be the solution to advanced control of future water treatment facilities.

- **Amanda S. Hering**, Baylor University
  "Comparison of Linear and Nonlinear Dimension Reduction Techniques for Automated Process Monitoring of a Decentralized Wastewater Treatment Facility"
  Karon Kazor, Ryan W. Holloway, Tzahi Y. Cath, and Amanda S. Hering*

  Multivariate statistical methods for online process monitoring have been widely applied to chemical, biological, and engineered systems. While methods based on principal component analysis (PCA) are popular, more recently kernel PCA (KPCA) and locally linear embedding (LLE) have been utilized to better model nonlinear process data. Additionally, various forms of dynamic and adaptive monitoring schemes have been proposed to address time-varying features in these processes. In this analysis, we extend a common simulation study in order to account for autocorrelation and nonstationarity in process data and comprehensively compare the monitoring performances of static, dynamic, adaptive, and adaptive-dynamic versions of PCA, KPCA, and LLE. Furthermore, we evaluate a nonparametric method to set thresholds for monitoring statistics and compare results with the standard parametric approaches. We then apply these methods to real-world data collected from a decentralized wastewater treatment system during normal and abnormal operations. From the simulation study, adaptive-dynamic versions of all three methods generally improve results when the process is autocorrelated and nonstationary. In the case study, adaptive-dynamic versions of PCA, KPCA, and LLE all flag a strong system fault, but nonparametric thresholds considerably reduce the number of false alarms for all three methods under normal operating conditions.

- **Gabriel J. Odom**, Baylor University
  "Multi-State Multivariate Statistical Process Control"
  Gabriel J. Odom*, Kathryn B. Newhart, Tzahi Y. Cath, and Amanda S. Hering

  Even though principal component analysis (PCA) is not optimal for autocorrelated, nonlinear, and non stationary data, adaptive-dynamic PCA (AD-PCA) has been shown to do as well as or better than nonlinear dimension reduction methods in flagging outliers in such environments. In some engineered systems, additional designed features create a known multi-state switching scheme among multiple autocorrelated, non-linear, and non-stationary processes, and incorporating this additional known information into AD-PCA can further improve it. In simulation with one of three types of faults introduced, we compare accounting for the states versus ignoring them. We find that multi-state AD-PCA reduces the proportion of false alarms and reduces the average time to fault detection. Conversely, we also investigate the impact of assuming multiple states when only one exists, and find that as long as the number of observations is sufficient, the misspecification is not detrimental. We then apply multi-state AD-PCA to real-world data collected from a decentralized wastewater treatment system during normal and abnormal operations. Multi-state AD-PCA flags a strong system fault earlier and more consistently than its single-state competitor. Furthermore, accounting for the physical switching system does not increase the number of false alarms under normal operating conditions and may ultimately assist with fault attribution.

Thursday 11:45AM–1:15PM

- **Julie Novak**, IBM T. J. Watson Research Center
  "Revenue Assessment in Large-Scale Businesses"
  Julie Novak*, Stefa Etchegaray and Yasuo Amemiya

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

- **Beatriz Etchegaray Garcia**, IBM T. J. Watson Research Center
"Predicting Staffing Needs for Optimal Opportunity Pipeline Management"
Beatriz Etchegaray Garcia

  Sales transaction support centers reduce the amount of transaction work performed by sellers allowing them to spend more time with customers and increase revenue generation. Adequate staffing ensures that staff with appropriate skills is available to support incoming service requests. We describe a suite of analytical tools including hierarchical forecasting and request assignment to determine staffing requirements. These tools were deployed across multiple geographies and lines of business.

- **David Banks**, Duke University
"Agent Based Models in Business"
David Banks

  Agent-based models are ubiquitous, and are becoming especially important in forecasting customer behavior and demand. This talk describes some of those applications, and points up the need for new statistical theory that is designed for the special problems that arise when making inferences from agent-based models.

- **Volodymyr Serhiyenko**, Metabiota
"Stochastic Modeling of Infectious Diseases"
Volodymyr Serhiyenko*, Nita Madhav, Cathine Lam, Nicole Stephenson, Kierste Miller, Mike Gahan, Mark Galliva and Ben Oppenheim.

  Infectious disease outbreaks can cause significant health, economic, and societal disturbances. The erratic nature of these events presents challenges for government agencies, corporations, and risk managers. Metabiota uses stochastic modeling to develop comprehensive analytic tools to help our private-sector and government clients mitigate and control the risk of complex biological threats. Here we will talk about disease spread model that can be used to stochastically simulate spread and duration of epidemics, using novel human coronaviruses such as those causing Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS) as an example. We will also discuss how a country's preparedness to epidemics can be incorporated into stochastic models through its influence on epidemiological parameters such as the case-fatality ratio and basic reproductive number.

- **Vincent Raja Anthonisamy**, University of Guyana
"Reliability Modeling Incorporating Load Share and Frailty"
G Asha, Vincent Raja Anthonisamy*, Nalini Ravishanker

  The stochastic behavior of lifetimes of a two component system is often influenced primarily by the system structure and by the covariates shared by the components. Any meaningful attempt to model the lifetimes must take into consideration the factors affecting their stochastic behavior. In particular, for a load share system, we describe a reliability model incorporating both the load share dependence and the effect of observed and unobserved covariates. The model includes a bivariate Weibull (Lu (1989)) to characterize load share, a positive stable distribution to describe frailty, and incorporates effects of observed covariates. We investigate various interesting reliability properties of this model including cross ratio functions and conditional survivor functions. We implement profile maximum likelihood estimation of the model parameters and discuss model adequacy and selection. We illustrate our approach using a simulation study. For a real data situation, we demonstrate the superiority of the proposed model which incorporates both load share and frailty effects over competing models that incorporate just one of these effects. An attractive and computationally simple cross -validation technique is introduced to reconfirm the claim. We conclude with a summary and discussion.

- **Vladimir Pozdnyakov**, University of Connecticut
"Discretely Observed Brownian Motion Governed by a Telegraph Process: Estimation"
Vladimir Pozdnyakov* and Jun Yan

  A Brownian motion whose infinitesimal variance alternates according to a telegraph process is considered. This stochastic process can be employed to model variety of real-word situations (for example, animal movement analysis or stochastic volatility modeling in mathematical finance). The main result

is an estimation procedure for underlying model parameters when the Brownian Motion governed by a telegraph process is observed discretely. Since the sequence of observations is not Markov, the likelihood estimation is done via dynamic programming.

- **Sanguthevar Rajasekaran**, University of Connecticut
  "Big Data Analytics: Basic Algorithms"
  Sanguthevar Rajasekaran

  Big data get generated in every area of science and engineering. Efficient techniques are needed to process these data. Specifically, useful information has to be obtained from massive data sets. Information extracted from biological data can result in gene identification, diagnosis for diseases, drug design, etc. Market-data information can be used for custom-designed catalogues for customers, supermarket shelving, and so on. Weather prediction and protecting the environment from pollution are possible with the analysis of atmospheric data. In this talk we present some challenges existing in processing big data in various disciplines. We also provide an overview of some basic algorithmic techniques. In particular, we will summarize various data analytics and reduction techniques.

- **Erin Conlon**, University of Massachusetts-Amherst
  "Parallel Markov Chain Monte Carlo Methods for Bayesian Analysis of Big Data"
  Erin Conlon

  Recently, new parallel Markov chain Monte Carlo (MCMC) methods have been developed for massive data sets that are too large for traditional statistical analysis. These methods partition big data sets into subsets, and implement parallel Bayesian MCMC computation independently on the subsets. The posterior MCMC samples from the subsets are then joined to approximate the full data posterior distributions. Current strategies for combining the subset samples include averaging, weighted averaging and kernel smoothing approaches. Here, I will discuss our new method for combining subset MCMC samples that directly products the subset densities. While our method is applicable for both Gaussian and non-Gaussian posteriors, we show in simulation studies that our method outperforms existing methods when the posteriors are non-Gaussian.

- **Kamiar Rahnama Rad**, CUNY
  "Scalable and Robust Model Estimation and Assessment"
  Kamiar Rahnama Rad*, Arian Maleki, Tim Machado and Liam Paninski

  The complexity of models and the massive size of structured big data calls for computationally efficient and statistically robust algorithms that avoid overfitting and undue bias. In this talk I will show how to take advantage of the high dimensionality of contemporary data sets to innovate efficient statistical methodologies for model selection and assessment. I will also demonstrate the robustness and scalability of this approach by applying it to both real data from the spinal cord and the entorhinal cortex.

Abstracts of Contributed Papers

- Stephen Clarke, SABIC Innovative Plastics
  "Historical Data Analysis with Autocorrelation"
  Stephen Clarke

  Historical data sets often exhibit large correlations among potential predictors as well as large autocorrelations among the observations. This latter problem is specifically addressed in the proposed process. With potentially hundreds (or thousands) of predictor variables, initial variable selection is required. Several techniques are available including Generalized (Lasso) Regression and Principle Components. Strict adherence to low Variance Inflation Factors, computed using Least Squares, is used to minimize multicollinearity among the predictors. Then Mixed Model Methodology is used to model the error structure of the data using an autoregressive (lag 1) error component. A Mixed Model analysis of an anonymized case study yields a different model and optimal solution than Ordinary Least Squares, demonstrating the risk of relying on Ordinary Least Squares. Criteria for removing irrelevant 2-Factor Interactions (2FIs) are presented. Lastly, establishing acceptable bounds on predictor variables is important when building optimization recommendations as well as visualization of 2FIs. If the range of the data (with thousands of observations) is used, then rare extreme values can obscure the realistic improvement opportunities.

- **Caleb King**, Sandia National Laboratories
  "Comparison of Accelerated Life Test Plans Based on Exact Small-Sample Methodology and Asymptotic Large-Sample Methodology"
  Caleb King

  The majority of the statistical literature on optimal test planning for accelerated life testing use the asymptotic variance of the quantity of interest as the criterion for optimality. Implicit in this procedure is the assumption that the number of samples used during testing is large enough to justify the underlying theory. However, it is often the case that financial and material constraints severely limit the number of samples available for testing. In these small-sample settings, a concern is that the test plans suggested by asymptotic theory may not yield the minimum variance. This paper represents a first attempt to develop optimal accelerated life test plans based on the exact variance of a quantile estimator with small samples providing an alternative to test plans based on large-sample theory. The optimal location of design points and sample allocation is determined for lognormal and Weibull lifetime distributions with complete data. The resulting test plans are then compared to those suggested by large-sample methods. In addition, the optimal small-sample test plans are used to determine the total number of samples needed to achieve a desired bound on the variance of the quantile estimator.

- **Raymond R. Hill**, Air Force Institute of Technolgoy
  "Examining Potential Reductions in Wind Tunnel Testing Data Requirements"
  Raymond R. Hill*, Doulas A. Dillard, Darryl K. Ahner, Douglas C. Montgomery

  This research explores the application of Design of Experiments (DOE) techniques in routine wind tunnel testing to reduce overall data requirements and demonstrates that significant reduction without information loss is possible. In addition, this research also shows there is a limit to how small a data set can be before a DOE design fails to remain statistically equivalent to a large data matrix created via the OFAT method. Practical applications of DOE to wind tunnel testing have significantly decreased wind-on minutes and total data volume compared to traditional tests. Thus, the DOE process can be used throughout the development of a major flight system to conserve resources. In addition, the research presents information loss differences between four distinctly different DOE designs; Covering Array, Nested Face Centered Design, I-Optimal Design, and Latin Hypercube. The information loss due to three different small sample sizes is quantified for a legacy wind tunnel test provided by Arnold Engineering Development Center, the sponsoring organization.

- **Iulian Ilies**, Northeastern University
  "Multimodality in Healthcare Length of Stay and Readmission Time Distributions: Implications on Outlier Detection and Prediction"
  Iulian Ilies*, James C. Benneyan

  Accurate prediction of inpatient stay durations is essential for effective patient flow management and utilization of limited resources such as specialty hospital beds. Prolonged stays contribute disproportionately to healthcare costs, increase adverse event risks such as infections and falls, and frequently are associated with early readmission rates. Correspondingly, a special emphasis has been placed on detection and prediction of high outlier lengths-of-stay and low outlier times- until-readmission. However, outlier identification criteria employed typically are fairly crude, ranging from expert-defined cutoff levels to simple functions of central tendency and dispersion measures. Underlying probability distributions are generally not considered beyond adjusting for right-skew, despite accumulating evidence of frequent multimodality. In the present study, we performed a systematic assessment of the impact of multimodality patterns on outlier detection in a large empirical dataset comprising over 155,000 admissions, including 50,000 readmissions. We stratified patients by diagnosis group, then optimally fitted mixtures of up to 5 right-skewed unimodal distributions to the corresponding length of stay and readmission time data, separately for each group. Multimodal distributions provided better fits than unimodal ones in 45-75% of cases, depending on the type of data and family of distributions used. Subsequent investigation of the overlap between and high or low density modes, respectively, and outlier sets based on commonly-used threshold rules, revealed substantial discrepancies. These findings suggest that distribution-free outlier criteria are often at risk of misclassifying patients because they overlook natural subgroups, which can then negatively impact failure analysis, improvement efforts, and the development of high-accuracy predictive models.

- **Wayne B. Nelson**, Wayne Nelson Statistical Consulting
  "An Improved Stockroom Reorder System"
  Wayne B. Nelson

  For reordering a stockroom item, a common rule of thumb is to order a six-month supply. This wastes money. A better approach is to model the costs of reordering, of handling, of idle items (money) sitting on the shelf, and of stockouts and then find the best reorder level and reorder quantity that minimize total cost for stocking an item. The usual textbook reorder model is usually inadequate. Through a case study, this talk presents a better reorder model for stockrooms that was used at General Electric Corp Research & Development. Implementing the model resulted in reducing stockroom operating cost about $250,000 annually (1987 dollars). Better modeling involved:
  - Including ALL costs of ordering, handling, shelving, storage space, and paying for an order of an item.
  - Modeling of non-Poisson demand for items.
  - Specifying an allowed probability of stockout rather than a cost of stockout.

  Before implementing the new reorder system, we determined whether it would be worthwhile. To do this, we drew a stratified random sample of stockroom items and calculated the anticipated savings for each and then estimated the total savings (worthwhile) for the entire population of stockroom items. Payback of the cost of implementing the system took nine months.

- **Martin Bezener**, Stat-Ease
  "Open Source Tools for Design of Experiments "
  Martin Bezener, Pat Whitcomb

  General purpose software for design of experiments (DOE) has traditionally been available in a handful of commercial packages. Currently available open source DOE tools are limited in scope, and fragmented across several packages and languages. Many freely available codes exist, but are typically specific to one method or paper, are not easy to use, and often require substantial work to integrate with other programs. Despite the recent increase in demand for open source DOE tools, there has not been much progress from the academic community. In this talk, I will begin with a quick survey of existing open source DOE packages in various programming languages. I will also spend part of

the talk describing a new DOE package in the Python language, dexpy. I will work through a few examples, including both traditional and computer generated optimal designs. Other issues, such as the nuances of licensing will also be discussed. I will conclude with some promising directions for future work.

- **Angelo Sant Anna**, Federal University of Bahia
  "Monitoring the Low Nonconforming Proportion of High-Quality Process Based on Beta Distribution"
  Angelo Sant Anna

  The quality of products has become an important aspect for the world class manufacturing in market competition and the quality control techniques such as statistical process control are becoming known in many industries. The advances of technologies have allowed to collect routinely valuable information from processes which theirs are analyzed using quality control methods. The high-quality processes have low rate of nonconforming products with small sample size available and the traditional p-chart is used for monitoring this problem. In the literature are found several works to improve the binomial probability distribution with normal approximation used to p-charts control limits. We propose the new control chart for monitoring low nonconforming products based on Beta distribution which substantially improve the control limits and to minimize problems in monitoring data from high-quality processes. The comparative study is illustrated for the proposed scheme and a simulation study is conducted to show the Beta chart is significantly superior to the several p-control charts in overall performance. Although the Beta chart is discussed in detail which can be applied to many other charts, such as CUSUM and EWMA charts.

- **Evgenii Sovetkin**, RWTH University, Aachen, Germany
  "Electroluminescence Image Analysis and Suspicious Areas Detection"
  Evgenii Sovetkin* and Ansgar Steland

  In this work we consider several problems arising in quality control analysis of electroluminescence (EL) images of photovoltaics (PV) modules. The EL image technique is a useful tool for investigating the state of a PV module and allows us to look inside a module and to analyse the crystalline structure at high resolution. However, there is a lack of methods to employ the information provided by EL images in the analysis of large PV systems. We first consider several practical issues that arise in field studies, i.e. when images are taken under outdoor conditions and not in a lab. We discuss a new problem-specific procedure for automatic correction of rotation and perspective distortions, which to some extent employs statistical approaches such as robust regression; and a procedure for automatic detection of the module and its cell areas (by means of a modified version of the Hough Transform). Those techniques provide us with images of the PV module cells, intensity light of which are of the main interest in quality study. Secondly, we discuss a simple test statistics to screen large databases of EL image data aiming at the detection of malfunctioning cells. The asymptotics is established for a general class of random fields, covering the asymptotic distribution and the estimation of the spatial covariances. Lastly, we present a spatial version of the test which aims at detecting the position of a defect and discuss asymptotic properties. The finite sample properties of the procedure are studied by simulations.

- **Kim Vukovinsky**, Pfizer
  "Statistical Contributions to the Pharmaceutical Quality by Design Effort"
  Kim Vukovinsky

  The FDAs Quality Initiative for the 21st century along with International Conference on Harmonization (ICH) Quality guidance documents provide both motivation and a framework for Quality by Design (QbD) implementation in the pharmaceutical industry. Designing quality into processes and demonstrating safe and efficacious products from these high quality processes requires an understanding of pharmaceutical and manufacturing sciences along with decision making in the presence of variability. Statistical thinking, tools, and methods assist the pharmaceutical, analytical method, and chemical process developers in developing measurement systems, formulations, and processes. This presentation will provide background on pharmaceutical process development, define terminology relevant to QbD, and overview statistical contributions to the development process.

- **Ke Wang**, Pfizer
  "Chemical Process Development Experimental Case Studies with Innovative Statistical Tools"
  Ke Wang

  Statistical practice influences and enhances decision making in the context of pharmaceutical scientific, business, and regulatory constraints. This talk will share some background on statistical collaboration with chemistry and engineering colleagues at Pfizer in advancing chemical process understanding and product development. Selected case studies including statistical design, modeling and risk quantification, will showcase how statistical practice in combination with chemistry principles innovates and contributes to chemical process design, characterization, range setting and the success of process validation.

- **Brent Harrington**, Pfizer
  "Analytical Method Development Translating the Analytical Target Profile to the Sampling Strategy"
  Brent Harrington

  As a regulated industry, pharmaceutical products and processes are governed by compendia rules. The rules constitute minimal standards for developing and verifying a process has manufactured a safe and efficacious product for consumer consumption. Demonstrating safety and efficacy demands an understanding of the science of pharmaceutical manufacture which means making development decisions under the uncertainty of quality measurements against the rules germane to that measurement. This talk illustrates how establishing a criterion from key measurement attributes and multiple compendia rules is utilized to assess the ability to make sound risk-based decisions. The talk will overview the resulting sampling strategy which consists of determining the number of dosage units (k) to be prepared in a composite sample of which there may be a number of equivalent replicate (r) sample preparations. Utilizing the most restrictive criteria in current regulatory documentation, a maximum variability attributed to method repeatability is defined and serves as the acceptance criterion. A table of solutions for the number of dosage units per sample preparation (r) and number of replicate sample preparations (k) is presented for any ratio of sample preparation and dosage unit variability.

- **Abdel-Salam Gomaa**, Qatar University
  "Cardiac Surgery Performance Monitoring via Control Charts"
  Abdel-Salam Gomaa

  A new CUSUM chart to monitor surgical performance in which the scores are adjusted to reflect the pre-operative estimate of the surgical risk for each patient (Steiner et al., 2000). They compared the performance of their proposed chart to Lovegrove et al's (1997) chart under only sustained step shifts. In this research, the risk-adjusted control charts in monitoring surgical performance will be utilized and analyzing a real data set obtained from a single surgical hospital from HMC in Qatar. The researcher also compares the performance of Steiner et al's (2000) approach to that of Lovegrove et al's (1997) approach under drift shift and unsustained shift with various values of odds ratio using a simulation study. The simulation study shows that Steiner et al's (2000) approach has better performance than Lovegrove et al's (1997) approach for all types of shifts. Thus it is used in the analysis of the real data set.

Abstracts of Posters

- **Renjie Chen**, University of Connecticut
  "Static and Dynamic Statistical Models for Daily Diary Data"
  Renjie Chen*, Nalini Ravishanker, Ofer Harel

  Collecting daily diary data is becoming increasingly useful in health outcome studies. The data consists of information on continuous-valued or binary responses and predictors on each of several subjects over a length of time, say a few months. Predictors may either be subject specific, or time-varying, or both. Some of the covariates are fixed-effects while others are random, leading to the use of mixed-effects modeling. The possible temporal correlation in the data over time may make results from the usual mixed-effects model inefficient. In this work, we compare the usual static mixed-effects modeling with modeling under repeated measures assumption (responses are assumed to be correlated over time, but regression coefficients are static) as well as under a dynamic mixed model (regression coefficients are assumed to be time-varying random variables). We implement the model fitting using SAS procedures (Mixed, NLMixed, Glimmix) and R packages (lme4, nlme, INLA). We present results from extensive simulation studies as well as from a real data application to study the effects of the different model assumptions on estimation and prediction.

- **Michael S. Czahor**, Iowa State University
  "Small-Scale Wind Turbine Recurrence Event Modeling as a Function of Environmental and Operational Covariates from Supervisory Control and Data Acquisition Systems"
  Michael S. Czahor* and William Q. Meeker

  Small-scale wind turbines (SWTs) are becoming increasingly popular as reduced costs and a small footprint allow wind to be used across many applications. Unfortunately, multiple downtime events raise concerns about the reliability and availability of SWTs. SWTs are repairable systems that can return to an operational state after a downtime or repair event. When a SWT experiences multiple events over time, these are known as recurrent events. The reliability of SWTs is examined in this paper using data from 21 individual 100 kW wind turbines. SWTs periodically record dynamic covariate data in the form of a vector time series using supervisory control and data acquisition (SCADA) systems. One type of event experienced by SWTs is known as a 'service event', which is a time when a SWT is put into service mode for a repair or false alarm. We explore the cost of each 'service event' and propose methodologies to link dynamic covariate data to downtime costs to assist in quantifying the variation of downtime across wind turbines. Data used in this work was provided from a power systems company in the United States. We outline a nonhomogenous Poisson process (NHPP) model with a Bayesian hierarchical power law structure in discrete-time that allows for inclusion of time-varying covariates.

- **Annesha Enam**, University of Connecticut
  "An Integrated Choice and Latent Variable Model for Multiple Discrete Continuous Choice Kernels: Application Exploring the Association Between Day Level Moods and Discretionary Activity Engagement Choices"
  Annesha Enam*, Karthik C. Konduri, Abdul R. Pinjari, Naveen Eluru

  In the recent years, multiple discrete continuous (MDC) model have emerged as popular framework to simultaneously model the choice of goods (that are imperfect substitutes to one another) and the associated consumption quantities. The poster presents a new integrated choice and latent variable (ICLV) model implementation called the Hybrid Multiple Discrete Continuous (HMDC) model that is capable of understanding the role of psychological factors (measured as latent constructs using indicator variables) on choice behaviors that can be represented using the MDC kernel. Estimation of ICLV models has been a challenge owing to the high dimensional integrals involved in the likelihood function. While maximum simulated likelihood estimation (MSLE) approach can be used, the methodology becomes cumbersome when the dimensionality of the integrals increases. In this research, a composite marginal likelihood (CML) based estimation approach is proposed for the estimation of the HMDC. Unlike the ICLV model implementations with single discrete choice kernel, the dimension of the integral to be decomposed in the HMDC varies across observations. This necessitated the use of weights when

decomposing the likelihood function using the CML approach. A simulation study was conducted using synthetic datasets to demonstrate the superiority of the weighted CML approach over its unweighted counterpart in the presence of MDC choice kernel. The applicability of the proposed model formulation and associated estimation routine was demonstrated using an empirical case study with data from the 2013 American Time Use Survey (ATUS).The empirical study identifies interesting association between day level moods and discretionary activity participation decision.

- **Amirsaman Hamzeh & Mahsa Mardikoraem** , University of Wisconsin-Milwaukee
  "Maximum Entropy Newsvendor Models"
  Amirsaman Hamzeh*, Mahsa Mardikoraem*, Ehsan S. Soofi

  The newsvendor model is characterized by finding the optimal order quantity for a perishable product with an uncertain demand and fixed prices. This model has been applied in production and operations management extensively under various probability distributions for the demand. Recently some authors have used maximum entropy models for the distribution of demand.The defining property of the newsvendor model is that ordering above the demand results in a salvage cost and ordering below the demand results in lost sales. This property partitions the support of the demand distribution below and above the order quantity and provides local information constraints for developing the maximum entropy model. We use the maximum entropy with local moment constraints which includes the newsvendor property. This formulation gives change point demand distributions. Examples include two-piece demand distributions with uniform, exponential, or truncated normal for demand below the order quantity and exponential or truncated normal for demand above the order quantity.

- **Jaeo Han**, Virginia Tech
  "using Social Network Analysis to Examine How Students Develop their Social Capital in a Classroom"
  Jaeo Han*, Shyam Ranganathan, and Denise Simmons

  As the measurement of students behavioral and emotional engagement, social capital, which the resources accrued through social networks, is a critical component to academic success. Student social capital can be developed by face-to-face or impersonal interactions with peers and/or faculty in a classroom. The goal of this study is to describe how differently students access and use resources embedded in social networks through social network analysis (SNA). Based on a newly developed survey instrument, we measured student perception of social capital across three types of the following measurements: value, reciprocity, conduit of belonging. We distributed the survey to more than 700 undergraduate students in 11 different classrooms at three institutions and received more than 500 responses, yielding a total response rate of 76.41% (554/725). Correlation matrix among network measures (e.g., In-degree, Out-degree, In-closeness, Out-closeness, betweenness, and eigenvalue centrality) indicated the existence of 'bridges' well-connected students who provide less well-connected students access to the available social capital in the class. At the same time, most communication among students in the class took place among students in small cliques where all students were connected to each other. Based on this data, we build models that predict the relative importance of students in the learning networks (based on network measures) as a function of the communication characteristics (length of conversations, topics discussed etc.) and student and class characteristics (gender, class type etc.). We also build models that predict that social capital exchanged between students as a function of these characteristics. Preliminary results suggest that the diversity of topics discussed and frequency of communication are significant predictors of the value of the communication (in terms of the social capital measures and network measures).

- **Yuan Jin**, University of Connecticut
  "Application of Finite Mixture Model in Crowdsourcing Quality Analysis"
  Sulin Ba, Yuan Jin*, Brian Lee, Jan Stallaert

  Contestants in a crowdsourcing contest can use the knowledge shared by others onto the platform to improve their performance and solution quality. But as contestants vary in their ways of using or learning from the shared knowledge, they could also be influenced by knowledge sharing differently. In this study, we apply a finite mixture model to identify contestants knowledge-using types and compare the effects of knowledge sharing on the performance and solution quality provided by contestants of

different types. We find that the finite mixture model fits our data better than a pooled regression, and provides more managerial implications for crowdsourcing and knowledge sharing management.

- **Eric Mittman**, Iowa State University
  "Baysian Hierarchical Model for Hard Drive Lifetime Data with Masked Failure Modes"
  Eric Mittman*, Colin Lewis-Beck

  The generalized limited failure population (GLFP) model, introduced by Chan and Meeker, is applicable to failure data where an unknown number of units are susceptible to early failure and all units are susceptible to wearout. Estimation of this model can encounter difficulties when few failures are observed and/or when left-truncation is present. We consider the case where group-level inference is desired for many groups with unequal sample sizes. In this setting, we avoid identifiability issues by assuming a common distribution across groups for early failures. Our Bayesian hierarchical model for grouped failure data uses data-driven shrinkage, making it possible to acheive a good model fit while accounting for uncertainty in all quantities of interest. We use the rstan package in R to fit the model to failure data on 75,000 hard-drives from various manufacturers made publicly available by Backblaze, a cloud storage company.

- **Shariq Mohammed**, University of Connecticut
  "Bayesian Variable Selection in High Dimensional Spatial Model with Application to Eeg Data using Spike and Slab Prior"
  Shariq Mohammed*, Dipak Dey and Yuping Zhang

  Due to the immense technological advances, very often we encounter data in high-dimensions. Any set of measurements taken at multiple time points for multiple subjects leads to data of more than two dimensions (matrix of covariates for each subject). In this poster, we present a Bayesian method for binary classification of subject-level responses by building binary regression models using latent variables along with the well known spike and slab priors. We also study the scaled normal priors on the parameters, as they cover a large family of distributions. Due to the computational complexity, we build many local (at different time points) models and aggregate the results. We do variable selection for each of these local models. If the variables are locations, then the variable selection can be interpreted as spatial clustering. We show the results of a simulation study and also present the performance of these models on multi-subject neuroimaging (EEG) data.

- **Kathryn Newhart**, Colorado School of Mines
  "Use of Principal Component Analysis for Early-Fault Detection in a Pilot-Scale Biological Wastewater Treatment System"
  Newhart, K.B.*, Odom, G.J., Hering, A.S., Cath, T.Y.

  Water scarcity is a stark reality for billions of people across the globe. Increasingly affluent populations in arid regions are putting stress on water resources, and solutions such as minimizing wasteful consumption and implementing water-efficient technologies are insufficient to meet the growing demand. Water reuse offers an opportunity to tap into an unconventional water source: highly-treated wastewater. Versatility in the water treatment process allows for water quality to be 'tailored' for potable or non-potable reuse; offsetting the burden on existing drinking-water sources and infrastructure. However, concern over effluent quality and the need for highly-skilled operators are two factors that have discouraged investment in water reuse. To overcome these hurdles, advanced and automated process control is being developed to detect and respond to system faults, to assure quality effluent at a low-cost. Early-fault detection allows preventative measures to be taken before water/wastewater treatment processes are compromised. For example, a change in influent quality or a minor system fault can result in a microbiological shift that disrupts proper treatment for months on small systems. Therefore, it is imperative to develop new water monitoring and treatment control methods as water reuse gains popularity. Current water/wastewater treatment process control is dominated by an operator's knowledge-base and pre-determined set-points. This approach to data-interpretation is limited in both scope and application. Due to the sheer volume and variety of data, it is impossible to process the data manually with the attention to detail required to understand fundamental relationships between water quality parameters and how changes in those parameters affect final effluent quality. Multivariate statistical analysis is able to identify abnormal relationships between a large number of

variables, despite their non-linearity. The goal of this research is to use an adaptive-dynamic statistical modeling approach to develop a smart, simple wastewater treatment program capable of detecting system faults in biological treatment and produce waters of different qualities for reuse applications. A demonstration-scale sequencing-batch membrane bioreactor (SB-MBR) is used to test the early-warning fault detection system. SB-MBR data was averaged over 10-minute intervals from when a fault was known to have occurred over the course of several days: the pH of the bioreactors slowly declined while the salinity of the influent increased. By the time the pH had fallen below an operator-defined set-point, the microbial community responsible for nitrogen removal was compromised and took months to recover. Had the early-warning system been online, the drop in pH would have been detected two days prior to the SB-MBR's system alarm. Further testing and development of the early-warning program includes online-monitoring of induced perturbations to enhance the fault-detection capabilities of the system and to train the program to respond to known or common faults. Additionally, the results will provide an in-depth analysis of individual water quality parameters. Hidden in this 'mountain' of data is information about how environmental and operational parameters can impact constituent removal in a treatment train, which could provide crucial, system-specific information without the need for advanced process or microbial modeling at water/wastewater treatment facilities worldwide.

- **Li Xu**, Virginia Tech
  "Statistical Methods for Prediction of High-Performance Computing i/o Variability"
  Li Xu*, Bo Li, Thomas Lux, Bo Li, Yili Hong, Layne Watson, and Kirk Cameron

  Managing performance variability is an important issue in high-performance computing (HPC). The performance variability is affected by complicated interactions of numerous factors, such as CPU frequency, the number of I/O threads, and I/O scheduler. In this paper, we develop statistical methods for modeling and analyzing HPC I/O variability. The objective is to identify both cause and magnitude of variability in HPC, and to predict performance variability. A complete analysis procedure is applied to deal with complex datasets. Several approximation methods are used to build predictive surface for the variability of HPC system. We evaluate the performance of the proposed method by leave-one-out test and mean prediction error in new system configurations. We also discuss the methodology for future system configuration by using tools in experimental designs.

- **Hong Yan**, Worcester Polytechnic Institute
  "Dynamic Space-Time Model for Syndromic Surveillance with Particle Filters and Dirichlet Process"
  Hong Yan, Zhongqiang Zhang, Jian Zou

  Massive spatio-temporal data are challenging for statistical analysis due to their low signal-to-noise ratios and high-dimensional spatio-temporal structure. To resolve these issues, we propose a novel Dirichlet process particle filter (DPPF) model. The Dirichlet process models a set of stochastic functions as probability distributions for dimension reduction, and the particle filter is used to solve the nonlinear filtering problem with sequential Monte Carlo steps where the data has a low signal-to-noise ratio. Our data set is derived from surveillance data on emergency visits for influenza-like and respiratory illness (from 2008 to 2010) from the Indiana Public Health Emergency Surveillance System. The DPPF develops a dynamic data-driven applications system (DDDAS) methodology for disease outbreak detection. Numerical results show that our model significantly improves the outbreak detection performance in real data analysis.

- **Yuan Yu**, Worcester Polytechnic Institute
  "Bayesian Bootstrap in Dealing with Sampling Bias with Application in Audit Data"
  Yuan Yu

  In auditing, the company may overstate their book values which will cause an error. Usually only a sample of accounts will be verified to estimate the error of the total population accounts. Since there are many zero errors in the audit data, previous research of Higgins and Nandram shows that, by introducing zero-inflation Poisson distribution to deal with the account errors population perform better to handle the real populations of account errors. In our study, we bring in Bayesian bootstrap sampling method for the poststratified errors which still leads to the similar estimation results while in

a non-parametric way. In the future, we also interest in sampling the unknown selection probabilities through a Bayesian non-parametric approach like Dirichlet Process.

- **Danilo Marcondes Filho**, Federal University of Rio Grande do Sul
  "Monitoring Batch Processes with Missing Variables"
  Danilo Marcondes Filho* and Luiz Paulo Luna de Oliveira

  Approaches to batch monitoring are mostly grounded on Multiway Principal Component Analysis (MPCA) due to the complexity of the underlying processes. Given the variety of scenarios, discussions and improvements about MPCA are found in the literature. However, there is a lack of works discussing batch monitoring in the important scenario where the set of available variables is incomplete. In fact, when some relevant variable is missing, such lack of information can substantially compromise MPCA modelling and diminish its ability to detect abnormal behaviors of future batch samples. Here we present an approach to deal with that problem proposing what we call T-MPCA, a MPCA modification by using the reconstruction method based on Takens theorem. Through this method some information of the missing variables can be recovered from correlating the measured variables with itself lagged in time. Besides its presentation, the efficacy of the new approach is illustrated using simulated data.

- **Angelo Sant Ana**, Federal University of Bahia, Salvador
  "Process Monitoring for Manufacturing data based on Model-based approach"

  Solving problems in industry, even inside companies known as expert in their sector, is not just a question of applying the right technique. The control chart is a traditional tool for data monitoring processes and the model-based approach has been shown to be very effective in detecting disturbances in output variables when input variables are measurable. The idea of the model-based control chart is to integrate generalized linear model and control chart tools to monitor any changes in process data. In many situations, there are variables that are nonconforming data following a Binomial distribution, and the modeling and monitoring this type data suffers serious inaccuracies in control limits specification when the rate of nonconforming is small. We propose the monitoring of nonconforming data using a GLM-based approach to the Binomial and the Beta model. Our purpose is to overcome such inaccuracies and discuss the performance of the proposed method in process monitoring. A case study is illustrated for the proposed method and a simulation study is conducted to compare the results using both the Binomial and Beta model-based charts. The paper presents also the structured steps to guide practitioners in the implementation of the GLM-based control charts for industrial processes monitoring.

- **Marco Bornstein and Ragib Mostofa**, Rice University
  "Applying the Efficient Market Hypothesis to Sports Binaries"
  Sathya Ramesh, Ragib Mostofa* and Marco Bornstein*

  The size of the global sports gambling market is estimated to be somewhere between $700 billion and $1 trillion. It would be highly lucrative to have a crystal ball to see into the future, and correctly bet on the outcomes of sporting events. Barring such discovery of the aforementioned magical technology, finding disparities between the market and reality to exploit market pricing inefficiencies and accurately predicting the results of sporting events is a potential method of generating profits in the unpredictable world of sports. Over the course of our research, we hoped to create an algorithmic system that would utilize data analysis and statistical techniques to forecast outcomes in order to generate positive returns for wagers made on National Basketball Association (NBA) and National Football League (NFL) games.

- **Philipp Wittenberg**, Helmut Schmidt University, Germany
  "A simple signaling rule for variable life-adjusted display derived from an equivalent risk-adjusted CUSUM chart"
  Philipp Wittenberg, Fah Fatt Gan, Sven Knoth

  The variable life-adjusted display (VLAD) is the first risk-adjusted graphical procedure proposed in the literature for monitoring the performance of a surgeon. It displays the difference between the cumulative sum of expected minus observed deaths. It has since become highly popular because the

statistic plotted is easy to understand. But it is also easy to misinterpret a surgeon's performance by utilizing the VLAD, potentially leading to grave consequences. The problem of misinterpretation is essentially caused by the VLAD's statistic variance that increases with sample size. In order for the VLAD to be truly useful, a simple signaling rule is desperately needed. In doing so, various forms of signaling rules have been developed but they are usually quite complicated. Without signaling rules, making inferences using the VLAD alone is difficult if not misleading. In this research project, we establish an equivalence between a VLAD with V-mask and a risk-adjusted cumulative sum (RA-CUSUM) based on the difference between the estimated probability of death and surgical outcome. Average run length analysis based on simulation shows that this particular RA-CUSUM chart has similar performance as compared to the established RA-CUSUM chart based on log-likelihood ratio statistic obtained by testing the odds ratio of death. We provide a simple design procedure for determining the V-mask parameters based on a resampling approach. Resampling from a real data set ensures that these parameters can be estimated appropriately. Finally, we illustrate the monitoring of a real surgeon's performance using VLAD with V-mask.

QPRC Technical Tour
Wednesday June 14, 2017
3:30 – 5:00 pm

We offer three exciting alternatives for the technical tour on the UConn Storrs campus. You may sign up for a tour at the registration desk on June 13. Space is limited on these tours, so please be flexible!

1. **Tour Option 1: Brain Imaging Research Center (BIRC). http://birc.uconn.edu/**
   Jay Rueckl – Director of BIRC / Associate Professor, Psychological Sciences

   BIRC is located in the David C. Phillips Communication Sciences Building on the Storrs campus. The 3,200 square foot BIRC provides a world-class environment for conducting research in the Cognitive Neurosciences, housing
   - 3 Tesla Siemens Prisma MRI Scanner
   - EEG (Electrical Geodesics, Inc.) with 256 channels and Eye Tracking (SR-Research Eye Link 1000 Plus)
   - MagVenture TMS (X100) integrated for simultaneous use with EEG and Eye Tracking systems.
   - Multiple behavioral testing rooms, a sound-attenuated testing chamber, and a high-end data processing and analytics lab, etc.

2. **Tour Option 2: Center for Hardware Assurance, Security and Engineering (CHASE) https://chase.uconn.edu/**
   John A. Chandy: Interim Director/Professor, ECE Department.

   The CHASE center was established in 2012 to provide the University with a physical and intellectual environment necessary for interdisciplinary hardware-oriented research and applications to meet the challenges of the future in the field of assurance and security. Among many other things, it houses
   - Keyence VHX2000: Super resolution Digital Microscope
   - Xradia Versa 510: State-of–the–Art X-ray Tomography with Nano Resolution
   - MPS150 Probe System: 150MM Manual Probing Solution for Failure Analysis
   - TERA OSCAT High-speed THz Spectrometer: Tetraherz Spectrometer

3. **Tour Option 3: Institute of Material Sciences (IMS). http://www.ims.uconn.edu/**
   Steven L. Suib - Director, IMS / Board of Trustees Distinguished Professor, Chemistry.

   Established in 1965, IMS operates and maintains extensive state-of-the-art instrumentation including a wide range of laboratories. It houses a wide range of advanced research instruments and facilities including
   - Atomic force and electron (SEM & TEM) microscopy
   - IR, Raman, UV, and laser spectroscopy
   - Nanobionics device fabrication facility, clean room
   - X-ray diffraction & spectroscopy (EDS, WDS, XRD, SAXS), etc.

Scholarship Recipients

Natrella Award Recipients

- Simon Tsz Fung Mak, Georgia Tech.
- Yan Wang, University of California, Los Angeles

QPRC Scholarship Recipients

- Evgenii Sovetkin, RWTH University, Aachen, Germany

NSF Scholarship Recipients

- Patrick Adams, University of Connecticut
- Atilla Ay, George Washington University
- Amirsaman Hamzeh Bajgiran, University of Wisconsin-Milwaukee
- Sudeep R. Bapat, University of Connecticut
- Marco Bornstein, Rice University
- Renjie Chen, University of Connecticut
- Patrick Christopher Daigle, University of Connecticut
- Michael S. Czahor, Iowa State University
- Annesha Enam, University of Connecticut
- Harshitha Girithara Gopalan , University of Louisville
- Jaeo Han, Institute of Technology
- Yuan Jin, University of Connecticut
- Jesse Kalinowski, University of Connecticut
- Michael Krause, American University
- Siddhesh Kulkarni, University of Connecticut
- Gemei Li, American University
- Yijun Liu, Fordham University
- Mahsa Mardikoraem, University of Wisconsin-Milwaukee
- Disheng Mao, University of Connecticut
- Eric Mittman, Iowa State University
- Shariq Mohammed, University of Connecticut
- Ragib Mostofa, Rice University
- Kathryn Newhart, Colorado School of Mines
- Gabriel J. Odom, Baylor University
- Nianting Ouyang, Fordham University
- Aaron Palmer, University of Connecticut

- Saad Quader, University of Connecticut
- Pedro Luiz Ramos, University of Connecticut
- Elham Sherafat, University of Connecticut
- Yulia Sidi, University of Connecticut
- Rui Sun, University of Connecticut
- Biju Wang, University of Connecticut
- Gang Wang, University of Connecticut
- Han Wang, Fordham University
- Li Xu, Virginia Institute of Technology
- Hong Yan, Worcester Polytechnic Institute
- Yuan Yu, Worcester Polytechnic Institute
- Yaohua Zhang, University of Connecticut
- Zhonghui Zhang, University of Connecticut
- Yuanshuo Zhao, Georgia Tech.

Notes