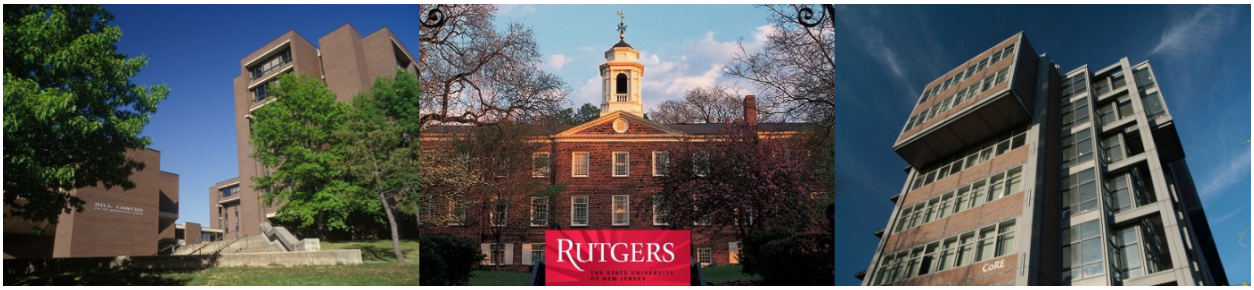

2017 Spring Research Conference

May 17-19, 2017

Rutgers, the State University of New Jersey
New Brunswick, NJ

DIMACS
Center for Discrete Mathematics and Theoretical Computer Science

WILEY

Contents

Sponsors

Spring Research Conference 2017 acknowledges the generous support from the following sponsors (in alphabetical order):

- American Statistical Association (ASA)
- Center for Discrete Mathematics and Theoretical Computer Science (DIMACS)
- Institute of Mathematical Statistics (IMS)
- International Chinese Statistical Association (ICSA)
- Procter & Gamble
- Rutgers, the State University of New Jersey
- Wiley

Committees

SRC management committee

Peter Qian, chair, U of Wisconsin-Madison
Xinwei Deng, Virginia Tech
Robert Gramacy, Virginia Tech
Ben Haaland, Georgia Tech
Pritam Ranjan, Acadiau University

SRC 2017 Organizing Committee:

Susan L. Albin (ISE), Yasuo Amemiya (IBM), Harry Crane, Elsayed A. Elsayed (ISE), Weihong Guo (ISE), Ying Hung (co-chair), Myong-Kee Jeong (ISE), John Kolassa, Regina Liu, William E Strawderman, Han Xiao, Min-ge Xie (co-chair), Dan Yang

General Information

KEYNOTE SPEAKER BIOGRAPHIES

David Madigan is the Executive Vice-President for Arts & Sciences, Dean of the Faculty, and Professor of Statistics at Columbia University in the City of New York. He previously served as Chair of the Department of Statistics at Columbia University (2008-2013), Dean, Physical and Mathematical Sciences, Rutgers University (2005-2007), Director, Institute of Biostatistics, Rutgers University (2003-2004), and Professor, Department of Statistics, Rutgers University (2001-2007). Dr. Madigan has over 160 publications in such areas as Bayesian statistics, text mining, Monte Carlo methods, pharmacovigilance and probabilistic graphical models. In recent years, he has focused on statistical methodology for generating reliable evidence from large-scale healthcare data. Dr. Madigan is a fellow of the American Association of the Advancement of Science (AAAS), the Institute of Mathematical Statistics (IMS) and the American Statistical Association (ASA), and an elected member of the International Statistical Institute (ISI). He served as Editor-in-Chief of *Statistical Science* (2008–2010) and *Statistical Analysis and Data Mining*, *the ASA Data Science Journal* (2013–2015).

Vijay Nair is currently Head of the Statistical Learning and Advanced Computing Group in Corporate Model Risk at Wells Fargo. He has been Donald A. Darling Professor of Statistics and Professor of Industrial & Operations Engineering at the University of Michigan since 1993. Previous to that, he was a Research Scientist at Bell Labs in NJ for 15 years. Vijay has broad research interests covering, among others, risk and reliability analysis, engineering statistics, design and analysis of experiments, and anomaly detection. Vijay has been editor of *Technometrics*, co-editor-in-chief of the *International Statistical Review*, and served on the editorial board of many other journals. He is a Fellow of AAAS, ASA, ASQ and IMS and has served as President of ISI and ISBIS. He has a PhD in Statistics from UC Berkeley.

BANQUET SPEAKER BIOGRAPHY

Xiao-Li Meng, Dean of the Harvard University Graduate School of Arts and Sciences (GSAS), Whipple V. N. Jones Professor and former chair of Statistics at Harvard, is well known for his depth and breadth in research, his innovation and passion in pedagogy, and his vision and effectiveness in administration, as well as for his engaging and entertaining style as a speaker and writer. Meng has received numerous awards and honors for the more than 150 publications he has authored in at least a dozen theoretical and methodological areas, as well as in areas of pedagogy and professional development; he has delivered more than 400 research presentations and public speeches on these topics, and he is the author of "The XL-Files," a regularly appearing column in the IMS (Institute of Mathematical Statistics) Bulletin. His interests range from the theoretical foundations of statistical inferences (e.g., the interplay among Bayesian, frequentist, and fiducial perspectives; quantify ignorance via invariance principles; multi-phase and multi-resolution inferences) to statistical methods and computation (e.g., posterior predictive p-value; EM algorithm; Markov chain Monte Carlo; bridge and path sampling) to applications in natural, social, and medical sciences and engineering (e.g., complex statistical modeling in astronomy and astrophysics, assessing disparity in mental health services, and quantifying statistical information in genetic studies). Meng received his BS in mathematics from Fudan University in 1982 and his PhD in statistics from Harvard in 1990. He was on the faculty of the University of Chicago from 1991 to 2001 before returning to Harvard as Professor of Statistics, where he was appointed department chair in 2004 and the Whipple V. N. Jones Professor in 2007. He was appointed GSAS Dean on August 15, 2012.

RECEPTION/POSTER SESSION

A reception and poster session will be held on Wednesday, May 17th, 5:15-6:45 p.m., on the first floor main lounge.

BANQUET

A banquet is scheduled for Thursday, May 18th, 6:30-8:30 p.m., and will take place in the multipurpose room.

List of Sessions

Wednesday, May 17

09:00-10:00	Keynote Address I	
10:30-12:00	Invited Session 1	*Computer Experiments I*
	Invited Session 2	*Challenges in Network Modeling and Sampling*
	Invited Session 3	*Data Fusion in Manufacturing*
13:30-15:00	Invited Session 4	*Statistical Methods for Uncertainty Quantification*
	Invited Session 5	*Journal of Quality Technology Session*
	Invited Session 6	*Power of Statistics in the Industry*
15:30-17:00	Invited Session 7	*Computer Experiments II*
	Invited Session 8	*Experimental Design and Causal Inference in High Tech Companies*
	Invited Session 9	*Modern Approaches to Multiple Testing, Multivariate data, and Missingness*
17:15–18:45	Reception/Poster Session	

Thursday, May 18

09:00-10:00	Keynote Address II	
10:30-12:00	Invited Session 10	*Design of Experiment I*
	Invited Session 11	*Predictive Modeling and Quality Control for Additive Manufacturing*
	Invited Session 12	*Time-series Analysis for Business Problems*
13:30-15:00	Invited Session 13	*Technometrics Session*
	Invited Session 14	*Reliability Modeling and Applications*
	Contributed Session 1	*New strategies for collection and*

		Analysis of data to solve modern-Day scientific and engineering Problems
15:30-17:00	Invited Session 15	*Design of Experiment II*
	Invited Session 16	*Recent Advances in Statistical Learning*
	Invited Session 17	*Machine Learning Algorithms in Complex Systems*
18:30 – 20:30	Banquet	

Friday, May 19		
08:30-10:00	Invited Session 18	*High Dimensional and Complex Time Series*
	Contributed Session 2	*Recent Advances in Design and Analysis of Classical Experiments*
	Contributed Session 3	*New Statistical and Machine Learning Methodologies Motivated by Cutting Edge Scientific Applications*
10:30-12:00	Invited Session 19	*Emerging Statistical Inference Methods in the Era of Data Science*
	Contributed Session 4	*New Developments in Space-Filling Designs*
	Contributed Session 5	*New Paradigms and Approaches in Modern-day Process Monitoring*

Conference Program Schedule

Day 1, Wednesday, May 17			
Time	**Multipurpose A**	**Multipurpose B**	**Room 411AB**
07:30 – 08:30	Continental Breakfast and Registration		
08:30 – 09:00	Welcome and Introductory Remarks		
09:00 – 10:00	Keynote Address I		
10:00 – 10:30	Coffee Break		
10:30 – 12:00	Invited Session 1 Computer Experiments I	Invited Session 2 Challenges in Network Modeling and Sampling	Invited Session 3 Data Fusion in Manufacturing
12:00 – 13:30	Lunch		
13:30 – 15:00	Invited Session 4 Statistical Methods for Uncertainty Quantification	Invited Session 5 Journal of Quality Technology Session	Invited Session 6 Power of Statistics in the Industry
15:00 – 15:30	Coffee Break		
15:30 – 17:00	Invited Session 7 Computer Experiments II	Invited Session 8 Experimental Design and Causal Inference in High Tech Companies	Invited Session 9 Modern Approaches to Multiple Testing, Multivariate data, and Missingness
17:15 – 18:45	Reception/Poster Session		

Wednesday, May 17, 8:30-9:00

Welcome and Introductory Remarks

Ying Hung, Department of Statistics and Biostatistics, Rutgers University

Ronald Ransome, Dean of Mathematics and Physical Sciences, Rutgers University

Tami Carpenter, Associate Director for the Center for Discrete Mathematics and Theoretical Computer Science (DIMACS), Rutgers University

Wednesday, May 17, 9:00-10:00

Plenary Session I - Chair: Peter Qian, University of Wisconsin-Madison

Statistical Methods for Risk Analysis in Banking
Vijay Nair, Wells Fargo & University of Michigan

Wednesday, May 17, 10:30-12:00

Invited Session 1: Computer Experiments I
Organizer: Matthew Pratola, Ohio State University
Chair: Bruce Ankenman, Northwestern University

Sequential Pareto Minimization of Physical System Means Using Calibrated Computer Simulators
Thomas Santner, Ohio State University

Spatial-temporal Kriging, Navier-stokes, and Combustion Instability
Jeff Wu, Georgia Institute of Technology

Invited Session 2: Challenges in Network Modeling and Sampling
Organizer and Chair: Harry Crane, Rutgers University

Edge Exchangeable Models for Interaction Networks
Walter Dempsey, University of Michigan

Random walk models of network formation
Peter Orbanz, Columbia University

Novel Approaches to Snowball Sampling That Circumvent the Critical Threshold
Karl Rohe, University of Wisconsin—Madison

Invited Session 3: Data Fusion in Manufacturing
Organizer and Chair: Ran Jin, Virginia Tech

Change-point Detection for Mixed-type Observations
Xinwei Deng, Virginia Tech

Random Planar Graphs to Quantify the Evolution of Surface Topography during Polishing Operations
Ashif Iquebal, Texas A&M University

Multi-sensors Data Fusion and Degradation Life Prediction
Changxi Wang, Rutgers University

Wednesday, May 17, 13:30-15:00

Invited Session 4: Statistical Methods for Uncertainty Quantification
Organizer and Chair: Qiong Zhang, Virginia Commonwealth University

An Empirical Adjustment of the Uncertainty Quantification in Gaussian Process Modeling
Daniel Apley, Northwestern University

Robust Parameter Design Using Computer Experiments
V. Roshan Joseph, Georgia Institute of Technology

Experimental Design Algorithms for Large-scale Statistical Computation
Peter Qian, University of Wisconsin-Madison

Invited Session 5: Journal of Quality Technology Session
Organizer: Fugee Tsung, Hong Kong University of Science & Technology
Chair: Bianca Colosimo, Politechico Di Milano, Italy

Multiple Objective Optimization in Reliability Demonstration Test
Lu Lu, University of South Florida

Planning Fatigue Tests for Polymer Composites
Yili Hong, Virginia Tech

Assessing Binary Measurement Systems with Targeted Verification
Daniel Severn, University of Waterloo

Invited Session 6: Power of Statistics in the Industry
Organizer and Chair: Shan Ba, the Procter & Gamble Company

Model Calibration with Censored Data
Shan Ba, the Procter & Gamble Company

Challenges of A/B Testing at Scale
Weitao Duan, LinkedIn Corporation

Forecasting with Big Data at Google
Weijie Shen, Google Inc.

Spatial Segmentation of Spatial-Temporal Lattice Models for Agricultural Management Zoning
Rodrigue Ngueyep Tzoumpe, IBM T.J. Watson Research Center

Wednesday, May 17, 15:30-17:00

Invited Session 7: Computer experiments II
Organizer: Bobby Gramacy, Virginia Tech
Chair: Christine Anderson-Cook, Los Alamos National Laboratory

Practical Heteroskedastic Gaussian Process Modeling for Large Simulation Experiments
Mickael Binois, University of Chicago Booth

A Variational Bayesian Inference-based Heteroscedastic Gaussian Process Approach for Simulation Metamodeling
Xi Chen, Virginia Tech

Modeling Heteroscedasticity with Bayesian Multiplicative Regression Trees
Hugh Chipman, Acadia University

Invited Session 8: Experimental Design and Causal Inference in High Tech Companies
Organizer and Chair: Tirthankar Dasgupta, Rutgers University

Randomized Experiments on Amazon's Supply Chain
David Afshartous, Amazon

Causal Inference at Google
Valeria Espinosa, Google

Ranking and Experimentation on LinkedIn Feed
Souvik Ghosh, LinkedIn

Trustworthy Analysis of Online A/B tests: Pitfalls, Challenges and Solutions
Jiannan Lu, Microsoft Corporation

Invited Session 9: Modern Approaches to Multiple Testing, Multivariate Data, and Missingness
Organizer and Chair: Arne Bathke, Salzburg University

Finding the Needle in the Haystack: Can Multivariate and Multiple Inference Methods Help?
Arne Bathke, Salzburg University

Inference for Related Sample When Some Information Is Missing
Solomon Harrar, University of Kentucky

Rank-Based Procedures in Factorial Designs: Hypotheses about Nonparametric Treatment Effects
Frank Konietschke, the University of Texas at Dallas

Wednesday May 17, 17:15-18:45

Reception/Poster Session

Day 2, Thursday, May 18			
Time	**Multipurpose A**	**Multipurpose B**	**Room 411AB**
07:30 – 09:00	Continental Breakfast and Registration		
09:00 – 10:00	Keynote Address II		
10:00 – 10:30	Coffee Break		
10:30 – 12:00	Invited Session 10 Design of Experiment I	Invited Session 11 Predictive Modeling and Quality Control for Additive Manufacturing	Invited Session 12 Time-series Analysis for Business Problems
12:00 – 13:30	Lunch		
13:30 – 15:00	Invited Session 13 Technometrics Session	Invited Session 14 Reliability Modeling and Applications	Contributed Session 1 New Strategies for Collection and Analysis of Data to Solve Modern-day Scientific and Engineering Problems
15:00 – 15:30	Coffee Break		
15:30 – 17:00	Invited Session 15 Design of Experiment II	Invited Session 16 Recent Advances in Statistical Learning	Invited Session 17 Machine Learning in Complex Systems
18:30 – 20:30	Banquet		

Thursday, May 18, 9:00-10:00

Plenary Session II - Chair: Regina Liu, Rutgers University

Honest Learning for the Healthcare System: Large-scale Evidence from Real-world Data
David Madigan, Columbia University

Thursday, May 18, 10:30-12:00

Invited Session 10: Design of experiments I
Organizer and Chair: Abhyuday Mandal, University of Georgia

Screening Designs, Potential Terms, Weighting, and Efficiency
JP Morgan, Virginia Tech

Thoughts on Optimal Statistical Design for Environmental Risk Assessment
John Stufken, Arizona State University

Strong Orthogonal Arrays of Strength Two Plus
Boxin Tang, Simon Fraser University

Invited Session 11: Predictive Modeling and Quality Control for Additive Manufacturing
Organizer and Chair: Arman Sabbaghi, Purdue University

Statistical Quality Monitoring in Metal Additive Manufacturing
Bianca Colosimo, Politechico Di Milano, Italy

Statistical Learning Issues in Quality Control for Additive Manufacturing
Qiang Huang, University of Southern California

Predictive Model Building Across Different Process Conditions and Shapes in 3D Printing
Arman Sabbaghi, Purdue University

Invited Session 12: Time-series analysis for business problems
Organizer: Yasuo Amemiya, IBM Watson Research Center
Chair: Rodrigue Ngueyep, IBM Watson Research Center

Case Sizes and the Bullwhip Effect
Chaitra Nagaraja, Fordham University

Statistical Challenges in Forecasting Revenue for Hierarchically Structured Businesses
Julie Novak, IBM T.J. Watson Research Center

Dynamic Models for Multivariate Times Series of Counts
Nalini Ravishanker, University of Connecticut

Thursday, May 18, 13:30-15:00

Invited Session 13: Technometrics Invited Session
Organize and Chair: Dan Apley, Northwestern University

Flexible Sequential or Sliced Designs for Computer Experiments
Mingyao Ai, Peking University

Optimization of Multi-fidelity Computer Experiments via the EQIE Criterion
Xu He, Chinese Academy of Sciences

Real-time Monitoring of High-Dimensional Functional Data Streams via Spatio-Temporal Smooth Sparse Decomposition
Kamran Paynabar, Georgia Institute of Technology

Invited Session 14: Reliability Modeling and Applications
Organizer and Chair: Elsayed, Rutgers University

Parallel Computing and Network Analytics for Fast Industrial Internet-of-Things (IIOT) Machine Information Processing and Condition Monitoring
Chen Kan, Penn State University

Analysis of Reliability Experiments with Blocking
Xiao Liu, University of Arkansas/IBM Thomas J. Watson Research Center

Learning from Multi-Modality Multi-Resolution Time Series Data: an Optimization Approach
Yada Zhu, IBM Research

Contributed Session 1：New Strategies for Collection and Analysis of Data to Solve Modern-day Scientific and Engineering Problems
Chair: John Kolassa, Rutgers University

Compromise Designs Under Baseline Parametrization
Ruwan Chamara Karunanayaka, Simon Fraser University

Using Particle Swarm Optimization to Search for Optimal Designs for Mixed Factor Experiments with Binary Response
Abhyuday Mandal, University of Georgia

Sequential Sampling Enhanced Composite Likelihood Approach to Estimation of Social Intercorrelations in Large-Scale Networks
Youran Qi, University of Wisconsin

Semi-parametric Adjustment to Computer Models
Rui Tuo, Georgia Institute of Technology

Thursday May 18, 15:30-17:00

Invited Session 15: Design of experiments II
Organizer and Chair: Frederick Phoa, Academia Sinica, Taiwan

Analyzing Definitive Screening Designs Taking Advantage of their Structure
Bradley Jones, JMP

The Construction of Space-Filling Designs with Good Uniformities in Multiple Dimensions via Factor Subgroup Collapse
Frederick Phoa, Academia Sinica, Taiwan

Closed-Loop Automatic Experimentation for Bayesian Optimisation
David Woods, University of Southampton

Invited Session 16: Recent Advances in Statistical Learning
Organizer: Cun-Hui Zhang, Rutgers University
Chair: Dan Yang, Rutgers University

Social Networks of Statisticians
Zheng Tracy Ke, The University of Chicago

Hypothesis Testing for Stochastic Block Models via Linear Spectral Statistics
Zongming Ma, University of Pennsylvania

Recent Developments in High-Dimensional Tensor Regression Problems
Garvesh Raskutti, University of Wisconsin

Stochastic Methods for Composite Optimization Problems
Feng Ruan, Stanford University

Invited Session 17: Machine Learning in Complex Systems
Organizer: Myong K (MK) Jeong and Grace Guo, Rutgers University
Chair: Grace Guo, Rutgers University

Comparison of Gaussian Process Modeling Software
Collin Erickson, Northwestern University

On Reject and Refine Options in Multicategory Classification
Xingye Qiao, Binghamton University

Physical-Statistical Modeling and Regularization of High-Dimensional Dynamic Systems
Bing Yao, Penn State University

A Nonparametric Approach for Partial Areas under ROC Curves
Yichuan Zhao, Georgia State University

Thursday May 18, 18:30-20:30

Banquet

Big Data, Big Surprises?
Xiao-Li Meng, Harvard University

Day 3, Friday May 19			
Time	**Multipurpose A**	**Multipurpose B**	**Room 411AB**
07:30 – 08:30	Continental Breakfast		
08:30 – 10:00	Invited Session 18 High Dimensional and Complex Time Series	Contributed Session 2 Recent Advances in Design and Analysis of Classical Experiments	Contributed Session 3 New Statistical and Machine Learning Methodologies Motivated by Cutting Edge Scientific Applications
10:00 – 10:30	Coffee Break		
10:30 – 12:00	Invited Session 19 Emerging Statistical Inference Methods in the Era of Data Science	Contributed Session 4 New Developments in Space-filling Designs	Contributed Session 5 New Paradigms and Approaches in Modern-Day Process Monitoring
12:30 – 13:00	Boxed Lunch		

Friday, May 19, 8:30-10:00

Invited Session 18: High Dimensional and Complex Time Series
Organizer and Chair: Han Xiao, Rutgers University

Regularized Estimation and Testing for High-Dimensional Multi-Block Vector Autorgressive Models
George Michailidis, University of Florida

Robust Testing and Variable Selection in High-Dimensional Time Series
Ruey Tsay, University of Chicago

Unsupervised Self-Normalized Change-Point Testing for Time Series
Ting Zhang, Boston University

Contributed Session 2: Recent Advances in Design and Analysis of Classical Experiments
Chair: David Woods, University of Southampton

Cmenet - a New Method for Bi-Level Variable Selection of Conditional Main Effects
Simon Mak, Georgia Institute of Technology

D-Optimal Mixture Designs for Ordinal Responses
Michelle Mancenido, Arizona State University

Supersaturated Designs Robust to Two-Factor Interactions
Chenlu Shi, Simon Fraser University

A Method for Identifying Dominant Effects in Designs with Complex Aliasing
Fasheng Sun, Northeast Normal University

Contributed Session 3: New Statistical and Machine Learning Methodologies Motivated by Cutting Edge Scientific Applications
Chair: Bianca Colosimo, Politechico Di Milano, Italy

Anisotropic Functional Laplace De-Convolution and Its Application to Dynamic Contrast Enhanced Imaging
Rida Benhaddou, Ohio University

The Blessing of Derivatives in Nonparametric Estimation
Xiaowu Dai, University of Wisconsin Madison

Bayesian Analysis of Traffic Flow Data
Vadim O. Sokolov, George Mason University

Generalized Support Vector Data Description with Bayesian Framework
Mehmet Turkoz, Rutgers University

Friday, May 19, 10:30-12:00

Invited Session 19: Emerging Statistical Inference Methods in the Era of Data Science
Organizer and Chair: Minge Xie, Rutgers University

On Integrative Learning of Mixed and Incomplete Data
Kun Chen, University of Connecticut

Recent Development on CPO Statistics in Joint Modeling of Longitudinal and Survival Data
Ming-Hui Chen, University of Connecticut

Testing for Homogeneity in Mixture Models
Yong Chen, University of Pennsylvania

Analysis Methods in the Presence of Mixed Measurement Error and Misclassification in Covariates
Grace Yi, University of Waterloo

Contributed Session 4: New Developments in Space-filling Designs
Chair: Boxin Tang, Simon Fraser University

Optimal Space-filling Latin Hypercube Designs Based on Good Lattice Point Sets
Lin Wang, UCLA

Space-filling Properties of Mirror-symmetric Orthogonal Designs
Yaping Wang, Peking University

Construction of Maximin Distance Latin Squares and Related Latin Hypercube Designs
Qian Xiao, UCLA

Construction of Nearly Uniform Designs on Irregular Regions
Jianfeng Yang, Nankai University

Contributed Session 5: New Paradigms and Approaches in Modern-Day Process Monitoring
Chair: Arman Sabbaghi, Purdue University

A Novel Pattern-Frequency Tree Approach for Transition Analysis and Anomaly Detection in Nonlinear and Nonstationary Systems
Cheng-Bang Chen, Penn State University

Sensor Fusion and On-line Monitoring of Friction Stir Blind Riveting for Lightweight Materials Manufacturing
Zhe Gao, Rutgers University

Nonparametric Change-Point Detection for Process Monitoring and Prognostics in Advanced Manufacturing
Shenghan Guo, Rutgers University

Statistical Process Monitoring of High-Dimensional Processes via Ridge
Sangahn Kim, Rutgers University

Abstracts

Wednesday, May 17

Plenary Session I

Statistical Methods for Risk Analysis in Banking
Vijay Nair, Wells Fargo & University of Michigan

Abstract Statistical modelling, analysis, and prediction have always been a major part of risk management in banking. This has become even more so in the regulatory environment after the financial crisis with "stress tests" to assess the viability and liquidity of large financial institutions. This presentation will: i) give an overview of risk modelling, ii) review different areas of risk, ranging from credit and market risk to financial crimes and operational risk, iii) discuss the use of statistical methods in a few applications, and iv) describe selected research problems.

Invited Session 1: Computer Experiments I

Sequential Pareto Minimization of Physical System Means Using Calibrated Computer Simulators
Thomas Santner, Ohio State University

Abstract The goal of this talk is to describe methodology for optimizing a manufacturing process when there are multiple, competing product objectives. This method assumes that both a simulator of the process is available as well as output from the manufacturing operation. The latter can be used to calibrate the simulator. The proposed approach identifies a set of manufacturing conditions each of which is on the Pareto Front of the product objectives, i.e., manufacturing conditions which cannot be modified to simultaneously improved all the product objectives. Sequential Designs based on the Minimax Fitness Function are used to efficiently add data from the simulator or from the manufacturing process, as the investigator desires. This method is illustrated with an example from an Injection Molding.
(joint work with PH Allen Chen, Angela Dean)

Spatial-temporal Kriging, Navier-stokes, and Combustion Instability
Jeff Wu, Georgia Institute of Technology

Abstract Most "learning" in big data is driven by the data alone. Some people may believe this is sufficient because of the sheer data size. If the physical world is involved, this approach is often insufficient. In this talk I will give a

recent study to illustrate how physics and data are used jointly to learn about the "truth" of the physical world. In the quest for advanced propulsion systems, a new design methodology is needed which combines engineering physics, computer simulations and statistical modeling. There are two key challenges: the simulation of high-fidelity spatial-temporal flows (using the Navier-Stokes equations) is computationally expensive, and the analysis and modeling of this data requires physical insights and statistical tools. First, a surrogate model is presented for efficient flow prediction in swirl injectors with varying geometries, devices commonly used in many engineering applications. The novelty lies in incorporating properties of the fluid flow as simplifying model assumptions, which allows for quick emulation in practical turnaround times, and also reveals interesting flow physics which can guide further investigations. Next, a flame transfer function framework is proposed for modeling unsteady heat release in a rocket injector. Such a model is useful not only for analyzing the stability of an injector design, but also identifies key physics which contribute to combustion instability.

Invited Session 2: Challenges in Network Modeling and Sampling

Edge Exchangeable Models for Interaction Networks
Walter Dempsey, University of Michigan

Abstract Many modern network datasets arise from processes of interactions in a population, such as phone calls, email exchanges, co-authorships, and professional collaborations. In such interaction networks, the edges comprise the fundamental statistical units, making a statistical framework for edge-labeled networks more appropriate for network analysis. In this context we introduce the class of edge exchangeable network models and develop their statistical properties. Edge exchangeable network models allow for sparse structure and power law degree distributions, both widely observed empirical properties that cannot be handled naturally by more conventional random graph models. The vertices in an edge exchangeable network arrive in size-biased order according to their degree, further explaining why vertex exchangeability is an untenable assumption for many applications.

Random walk models of network formation
Peter Orbanz, Columbia University

Abstract I will describe a class of network models that insert edges by connecting the starting and terminal vertices of a random walk on the network graph. Within the taxonomy of statistical network models, this class is distinguished by permitting the location of a new edge to explicitly depend on the structure of the graph, but being nonetheless statistically and computationally tractable. In the limit of infinite walk length, the model converges to an extension of the preferential attachment model. I will discuss theoretical properties, such as power laws, and show that inference of model

parameters is possible from a single graph generated by the model. (Joint work with Benjamin Bloem-Reddy.)

Novel Approaches to Snowball Sampling That Circumvent the Critical Threshold
Karl Rohe, University of Wisconsin—Madison

Abstract Web crawling, snowball sampling, and respondent-driven sampling (RDS) are three types of network driven sampling techniques that are popular when it is difficult to contact individuals in the population of interest. It has been previously shown that if participants refer "too many" other participants, then the standard approaches do not provide "square root n" consistent estimators. This talk will discuss two novel approaches to provide such estimators. The first approach must be incorporated during data collection. The second approach is a novel estimator. This talk will discuss both statistical theory and a human subject experiment to test the validity of the novel form of data collection.

Invited Session 3: Data Fusion in Manufacturing

Change-point Detection for Mixed-type Observations
Xinwei Deng, Virginia Tech

Abstract Data with mixed-type characteristics are ubiquitous in many applications. This talk mainly focuses on change-point detection problem for mixed-type observations. The main challenge lies on how to quantify the hidden association among mixed-type observations. We proposed a latent process method, so-called mixed switching state-space model, to jointly model mixed-type observations and effectively detect the change-points. Efficient parameter estimation and Bayesian inference are developed by iteratively combining discrete particle filter and sequential Monte Carlo algorithms. Both simulation and read data examples are used to elaborate the performance of the proposed method.

Random Planar Graphs to Quantify the Evolution of Surface Topography during Polishing Operations
Ashif Iquebal, Texas A&M University

Abstract Recent results on the spectral properties of planar graphs provide an immense potential to model the evolution of surface topography in precision manufacturing processes. The need for such models is growing with the increasing emphasis on post-processing stages, such as polishing, of additive manufactured components, especially for biomedical and aerospace applications. Currently, very limited efforts have been made towards the quantification of surface topography to facilitate real-time surface quality monitoring and end-point detection. Our recent investigations show that as

polishing ensues, the asperities progressively level with significant plastic flow of the materials in the form of thin fluid-like layers that subsequently bridge to result in a smooth finish. We present a random planar graph representation to model the topography of the surface where the nodes represent the individual asperities, and the edges represent the propensity of the neighboring asperities to form a bridge. We provide theoretical bounds on spectral quantifiers of the graphs by invoking the packing density of hard spheres that connote the various stages, including the end, of polishing. Comparative results using in situ electron micrographs collected during various stages of polishing suggest that spectral graph quantification presents an effective representation of local surface characteristics e.g., micro-pores and defects as compared to traditional surface roughness quantifiers.

Multi-sensors Data Fusion and Degradation Life Prediction
Changxi Wang, Rutgers University

Abstract The rapid developments of sensing technology and data acquisition systems has led to an ever-increasing interest in continuous monitoring of degradation data with multiple sensors. Since different sensors portray different aspects of the degradation process and each sensor data contains only partial information of the degraded component, data fusion approaches that integrate degradation data from multiple sensors can effectively improve the degradation life prediction accuracy. We present a new approach that assigns weights for every sensors observation based on dynamic clustering of the sensors observations with time. In order to improve data fusion accuracy in tracking the degradation path, a probabilistic clustering approach is used and the degradation path is compared with that obtained by k-means clustering approach. A case study that involves a simulated degradation dataset is implemented to numerically evaluate and compare the prognostic performance of the proposed method with that with others.

Invited Session 4: Statistical Methods for Uncertainty Quantification

An Empirical Adjustment of the Uncertainty Quantification in Gaussian Process Modeling
Daniel Apley, Northwestern University

Abstract Gaussian process (GP) models have emerged as the standard surrogate models for deterministic computer response surfaces, due in part to their built-in mechanism for providing uncertainty quantification (UQ) in the form of prediction intervals on the response, in addition to the response prediction itself. However, their covariance parameter estimation methods tend to favor giving good response prediction, at the expense of poor UQ. We propose and investigate a post-processing method that takes a fitted GP model and its training data, and then empirically adjusts the built-in GP UQ so that it is in better agreement with the actual uncertainty in the response predictions. We

demonstrate that this substantially improves the accuracy of the UQ, especially for nonstationary response surfaces that have more complex behavior is some input regions.

Robust Parameter Design Using Computer Experiments
V. Roshan Joseph, Georgia Institute of Technology

Abstract In this talk I will discuss new experimental design methods for computer simulations involving noise factors. Most existing methods focus on uniformly distributing the points in the design space, which are not suitable for noise factors because they usually follow non-uniform distributions such as normal distribution. This would suggest placing more points in the regions with high probability mass. However, noise factors also tend to have a smooth relationship with the response and therefore, placing more points towards the tails of the distribution is also useful for accurately estimating the relationship. These two opposing effects make the experimental design methodology a challenging problem. We propose optimal and computationally efficient solutions to this problem and demonstrate their advantages using simulated examples and a real industry example involving a manufacturing packing line. (Joint work with Li Gu, Shan Ba, and William Myers).

Experimental Design Algorithms for Large-scale Statistical Computation
Peter Qian, University of Wisconsin-Madison

Abstract Big Data appear in a growing number of areas like marketing, physics, biology, engineering, and the Internet. For example, for every hour, more than one million transaction data are stored in WalMart database and a HPC based computer model can produce results of millions of runs. While large volume of data offers more statistical power, it also brings computational challenges. We first review an experimental design algorithm, called orthogonalizing EM (OEM), intended for various least squares problems. The main idea of the procedure is to orthogonalize an arbitrary design matrix by adding new rows and then solve the original problem by embedding the augmented design in a missing data framework. We demonstrate that OEM is highly efficient for large-scale least squares problems.
We then present a reformulation and generalization of OEM that leads to a reduction in computational complexity for least squares and penalized least squares problems. The reformulation, named the GOEM (Generalized Orthogonalizing EM) algorithm, is further extended to a wider class of models including generalized linear models and Cox's proportional hazards model. Synthetic and real data examples are included to illustrate its efficiency compared with standard techniques. Inspired by the idea of OEM, we present an iterative algorithm, named iKriging, for fitting large-scale Gaussian process models for computer experiments.

Invited Session 5: Journal of Quality Technology Session

Multiple Objective Optimization in Reliability Demonstration Test
Lu Lu, University of South Florida

Abstract Reliability demonstration tests are commonly performed in product development or a validation process to demonstrate whether a product meets specified requirements on reliability. Among binomial demonstration tests, zero-failure tests have been most commonly used for their simplicity and the use of minimum sample size to achieve an acceptable consumer's risk level. However, these often result in unacceptably high risks for producers as well as low probabilities of passing the tests even when products have adequate reliability. In this talk, we explicitly explore the interrelationship between multiple objectives that are commonly of interest when planning a demonstration test, and propose structured decision-making strategies using a Pareto front based approach for selecting optimal test plans that simultaneously balancing multiple criteria. A variety of strategies are suggested for scenarios with different user priorities, and graphical tools are developed to effectively quantify the trade-offs between choices and to facilitate informed decision-making. Potential impacts on the final decision of some subjective user inputs are studied to offer insights and useful guidance for general applications. (Note this will be a joint presentation with Dr. C.M. Anderson-Cook from Los Alamos National Laboratory.)

Planning Fatigue Tests for Polymer Composites
Yili Hong, Virginia Tech

Abstract Polymer-composite materials have become key components in the transportation and alternative-energy industries, as they are more lightweight than homogeneous metals and alloys yet still retain comparable levels of strength and endurance. To understand how these polymer composites perform after long periods of use, material manufacturers commonly use cyclic fatigue testing. The current industrial standards include test plans with balanced designs and equal spacing of stress levels which, in many cases, are not the most statistically efficient designs. In this paper, we present optimal designs with the goal of minimizing the weighted sum of asymptotic variances of an estimated lifetime percentile at selected design stress levels. These designs are based on a physical model adapted from the fatigue literature, which is more suitable for modeling cyclic fatigue of polymer composites than the model used in the current industrial standards. We provide a comparison between our optimal designs and the traditional designs currently in use and ultimately propose a compromise design for use by practitioners in order to ensure robustness against deviations from the underlying assumptions. This is a joint work with Caleb King, Stephanie DeHart, Patrick DeFeo, and Rong Pan.

Assessing Binary Measurement Systems with Targeted Verification

Daniel Severn, University of Waterloo

Abstract This talk will discuss the concept of Targeted Verification and its implementation. It considers the situation where an imperfect non-destructive binary measurement system requires assessment when a gold standard measurement system is available. The context usually dictates that using the gold standard measurement system to verify parts is costly or otherwise burdensome to use. Targeted Verification is a design element of a sequential statistical plan that targets specific parts for verification based on the outcome of previous phases in the sequential plan. This plan dramatically reduces the number of parts that need to be verified while attaining performance similar to that of plans that verify all parts and avoiding the pitfalls of plans that verify no parts.

Invited Session 6: Power of Statistics in the Industry

Model Calibration with Censored Data
Shan Ba, The Procter & Gamble Company

Abstract The purpose of model calibration is to make the model predictions closer to reality. The classical Kennedy-O'Hagan approach is widely used for model calibration, which can account for the inadequacy of the computer model while simultaneously estimating the unknown calibration parameters. In many applications, the phenomenon of censoring occurs when the exact outcome of the physical experiment is not observed, but is only known to fall within a certain region. In such cases, the Kennedy-O'Hagan approach cannot be used directly, and we propose a method to incorporate the censoring information when performing model calibration. The method is applied to study the compression phenomenon of liquid inside a bottle. The results show significant improvement over the traditional calibration methods, especially when the number of censored observations is large.

Challenges of A/B Testing at Scale
Weitao Duan, LinkedIn Corporation

Abstract A/B testing (a.k.a. controlled experiments) is utilized to make data-driven decisions at many Internet companies. Statistics has always been the backbone of A/B testing, and its role is even more crucial in the era of Big Data. Over the past few years we have seen a massive growth of experimentation at LinkedIn. We will discuss several important statistical challenges when running A/B testing at Internet scale. Especially we will share our effort in running experiments under the complex network structure.

Forecasting with Big Data at Google
Weijie Shen, Google Inc.

Abstract The need for reliable forecasts at scale has brought up new challenges and opportunities at Google. From abundant data sources, time series come in both in a finer resolution and with a wider coverage. These features not only improve the quality of the forecasts but also help address the questions we were not able to decades ago. Meanwhile, they also introduce new challenges such as scalability, robustness and consistency. In this talk, I will briefly introduce the use of time series forecasting at Google, touch on some new opportunities and challenges, and some best practices. A case study of different forecast tools will be provided for illustration.

Spatial Segmentation of Spatial-Temporal Lattice Models for Agricultural Management Zoning
Rodrigue Ngueyep Tzoumpe, IBM T.J. Watson Research Center

Abstract In many applications where both predictors and responses are collected across geographical regions over time, the impact of the predictors to responses are often not static but time-varying. Moreover, the time-varying impact of the predictors may vary across different regions. To identify nearby regions where these time-varying impact behave similarly, we propose a spatially fused time-varying lattice model. We model time-varying impact of spatio-temporal predictors via a spatial lattice model with time-varying coefficients. Furthermore, we utilize fusion penalty to allow nearby regions to share same time-varying coefficients. The model parameters can be efficiently estimated via ADMM algorithm. One motivation application of our method is to identify agriculture management zones where the time-varying impact of environment attributes (e.g., growing degree days, heat stress, precipitation) on the crop yield is similar. Once these zones are identified, same planting policy could be implemented within these zones.

Invited Session 7: Computer experiments II

Practical Heteroskedastic Gaussian Process Modeling for Large Simulation Experiments
Mickael Binois, University of Chicago Booth

Abstract We present a unified view of likelihood based Gaussian progress regression for simulation experiments exhibiting input-dependent noise. Replication plays an important role in that context, however previous methods leveraging replicates have either ignored the computational savings that come from such design, or have short-cut full likelihood-based inference to remain tractable. Starting with homoskedastic processes, we show how multiple applications of a well-known Woodbury identity facilitate inference for all parameters under the likelihood (without approximation), bypassing the typical full-data sized calculations. We then borrow a latent-variable idea from machine learning to address heteroskedasticity, adapting it to work within the same thrifty inferential framework, thereby simultaneously leveraging the

computational and statistical efficiency of designs with replication. The result is an inferential scheme that can be characterized as single objective function, complete with closed form derivatives, for rapid library-based optimization. Illustrations are provided, including real-world simulation experiments from manufacturing and the management of epidemics.

A Variational Bayesian Inference-based Heteroscedastic Gaussian Process Approach for Simulation Metamodeling
Xi Chen, Virginia Tech

Abstract We propose a variational Bayesian inference-based Gaussian process metamodeling approach (VBGP) suitable for stochastic simulation experiments. This approach enables simultaneous approximations to the mean and variance response surfaces implied by a stochastic simulation, and it can accommodate situation where simulation replications are not available at every design point. The use of variational Bayesian inference enables more accurate mean response prediction while taking into account the uncertainty arising from estimation of heterogeneous simulation variances. We demonstrate the superior performance of VBGP through some numerical examples.

Modeling Heteroscedasticity with Bayesian Multiplicative Regression Trees
Hugh Chipman, Acadia University

Abstract Bayesian additive regression trees (BART) has become increasingly popular as a flexible and scalable non-parametric model useful in many modern applied statistics regression problems. It brings many advantages to the practitioner dealing with large and complex non-linear response surfaces, such as the matrix-free formulation and the lack of a requirement to specify a regression basis a priori. However, while flexible in fitting the mean, the basic BART model uses the standard i.i.d. normal model for the errors. This assumption is unrealistic in many applications. Moreover, in many applied problems understanding the relationship between the variance and predictors can be just as important as that of the mean model. We develop a novel heteroscedastic BART model to alleviate these concerns. Our approach is entirely non-parametric and does not rely on an a priori basis for the variance model. In BART, the conditional mean is modeled as a sum of trees, each of which determines a contribution to the overall mean. In this talk, we model the conditional standard deviation as the result of a product of trees, each of which determines a contribution to the overall standard deviation. The approach will be demonstrated using a combination of simulated and real datasets.

Invited Session 8: Experimental Design and Causal Inference in High Tech Companies

Randomized Experiments on Amazon's Supply Chain
David Afshartous, Amazon

Abstract Amazon leadership principles (LPs) include Earn Trust, Invent and Simplify, Learn and Be Curious, Insist on the Highest Standards, and Dive Deep, which are also tenets that lead to good science. Our tenacious adherence to our LPs fosters a superb research environment that has helped make us the number 1 e-commerce retailer in the world. At Inventory Planning and Control Laboratory (IPC Lab) we run randomized controlled trials (RCTs) that evaluate the efficacy of in-production buying and supply chain policies on important business metrics. Our customers are leading supply chain researchers and business managers within Amazon, and our mission is to help them best answer the question, 'Should I roll out my policy?' In this talk we discuss how our LPs guide us to overcoming multiple obstacles to fulfilling our mission, including scalability, non-Gaussian data, and domain-specific challenges. Our self-service platform allows our customers to hundreds experiments annually using data from our custom data store. Our experiments can contain hundreds of millions of observations, necessitating scalable statistical methodologies. Despite our large datasets, determining the correct data-driven decision is difficult, as our data often exhibit high proportions of zeros, extreme right-tailed skewness, and small beneficial treatment effects. Finally, we discuss issues of communication and translating experimental results to assessment of the financial impact.

Causal Inference at Google
Valeria Espinosa, Google

Abstract Experimentation is at the heart of Google. We will discuss several of the most impactful experimental and observational methods that Google has utilized and developed.

Ranking and Experimentation on LinkedIn Feed
Souvik Ghosh, LinkedIn

Abstract The LinkedIn feed for a member is built with the activities of her connection network. The member can take various actions on the feed such as liking or commenting on the activities. These activities then propagate through the network to many more members. Ranking feed updates for a member involve showing the most relevant updates at the top of the feed. In this talk we will discuss the algorithms used for ranking feed updates and the challenges in measuring the impact of such algorithms in online experiments.

Trustworthy Analysis of Online A/B Tests: Pitfalls, Challenges and Solutions
Jiannan Lu, Microsoft Corporation

Abstract A/B tests (or randomized controlled experiments) play an integral role in the research and development cycles of technology companies. As in classic randomized experiments (e.g., clinical trials), the underlying statistical analysis

of A/B tests is based on assuming the randomization unit is independent and identically distributed (i.i.d.). However, the randomization mechanisms utilized in online A/B tests can be quite complex and may render this assumption invalid. Analysis that unjustifiably relies on this assumption can yield untrustworthy results and lead to incorrect conclusions. Motivated by challenging problems arising from actual online experiments, we propose a new method of variance estimation that relies only on practically plausible assumptions, is directly applicable to a wide of range of randomization mechanisms, and can be implemented easily. We examine its performance and illustrate its advantages over two commonly used methods of variance estimation on both simulated and empirical datasets. Our results lead to a deeper understanding of the conditions under which the randomization unit can be treated as i.i.d. In particular, we show that for purposes of variance estimation, the randomization unit can be approximated as i.i.d. when the individual treatment effect variation is small; however, this approximation can lead to variance under-estimation when the individual treatment effect variation is large.

Invited Session 9: Modern Approaches to Multiple Testing, Multivariate Data, and Missingness

Finding the Needle in the Haystack: Can Multivariate and Multiple Inference Methods Help?
Arne Bathke, Salzburg University

Abstract When there are several relevant responses and different experimental conditions, researchers typically want to find out which conditions are relevant, and for which responses. We present two rather general approaches trying to accomplish these goals, accommodating binary, ordinal, and metric endpoints, and different nominal factors. We also try to address the question of how well the proposed methods actually accomplish their goals.

Inference for Related Sample When Some Information Is Missing
Solomon Harrar, University of Kentucky

Abstract In this talk, we present accurate inferential methods for related samples. The methods are designed for the situation where outcome is missing or information about group membership is subject to error. Parametric, semi-parametric and rank-based methods that are accurate for small to moderate samples are developed. The principal tools for getting accuracy in small to moderate samples are Edgeworth expansions and resampling techniques. Numerical studies are used to investigate the performance of methods. Real data examples from addiction intervention and medical imaging studies are used to illustrate the application of the methods.

Rank-Based Procedures in Factorial Designs: Hypotheses about Nonparametric Treatment Effects
Frank Konietschke, the University of Texas at Dallas

Abstract Existing tests for factorial designs in the nonparametric case are based on hypotheses formulated in terms of distribution functions. Typical null hypotheses, however, are formulated in terms of some parameters or effect measures, particularly in heteroscedastic settings. In this talk we extend this idea to nonparametric models by introducing a novel nonparametric ANOVA-type-statistic based on ranks which is suitable for testing hypotheses formulated in meaningful nonparametric treatment effects in general factorial designs. This is achieved by a careful in-depth study of the common distribution of rank-based estimators for the treatment effects. Since the statistic is asymptotically not a pivotal quantity we propose different approximation techniques, discuss their theoretic properties and compare them in extensive simulations together with two additional Wald-type tests. An extension of the presented idea to general repeated measures designs is briefly outlined. The proposed rank-based procedures maintain the pre-assigned type-I error rate quite accurately, also in unbalanced and heteroscedastic models. A real data example illustrates the application of the proposed methods

Thursday, May 18

Plenary Session II

Honest Learning for the Healthcare System: Large-Scale Evidence from Real-World Data
David Madigan, Columbia University

Abstract In practice, our learning healthcare system relies primarily on observational studies generating one effect estimate at a time using customized study designs with unknown operating characteristics and publishing -- or not -- one estimate at a time. When we investigate the distribution of estimates that this process has produced, we see clear evidence of its shortcomings, including an over-abundance of estimates where the confidence interval does not include one (i.e. statistically significant effects) and strong indicators of publication bias. In essence, published observational research represents unabashed data fishing. We propose a standardized process for performing observational research that can be evaluated, calibrated and applied at scale to generate a more reliable and complete evidence base than previously possible, fostering a truly learning healthcare system. We demonstrate this new paradigm by generating evidence about all pairwise comparisons of treatments for depression for a relevant set of health outcomes using four large US insurance claims databases. In total, we estimate 17,718 hazard ratios, each

using a comparative effectiveness study design and propensity score stratification on par with current state-of-the-art, albeit one-off, observational studies. Moreover, the process enables us to employ negative and positive controls to evaluate and calibrate estimates ensuring, for example, that the 95% confidence interval includes the true effect size approximately 95% of time. The result set consistently reflects current established knowledge where known, and its distribution shows no evidence of the faults of the current process. Doctors, regulators, and other medical decision makers can potentially improve patient-care by making well-informed decisions based on this evidence, and every treatment a patient receives becomes the basis for further evidence.
(joint work with Martijn J. Schuemie, Patrick B. Ryan, George Hripcsak, and Marc A. Suchard)

Invited Session 10: Design of Experiments I

Screening Designs, Potential Terms, Weighting, and Efficiency
JP Morgan, Virginia Tech

Abstract DuMouchel and Jones (1994) introduced an approach to D-optimal design selection allowing estimation of all primary terms (those in the assumed model) with good detectability for potential terms (terms which may or may not be contributors to the actual data generating mechanism). An asymptotic examination of their modified information matrix establishes a connection with a corresponding weighted information matrix. Combining the two techniques provides another avenue for generating useful fractional factorial designs. This approach is examined in terms of efficiency and potential for bias.

Thoughts on Optimal Statistical Design for Environmental Risk Assessment
John Stufken, Arizona State University

Abstract A problem that comes up in environmental risk studies is the estimation of a benchmark dose for polluting agents under a dose- response model. A complication is that there is a great deal of model uncertainty. For a study one needs to select doses for the pollutants at which measurements are to be made, which is a design problem. This design problem has not been studied in great detail. We describe the problem and propose a method that helps to identify efficient designs for estimation of the benchmark dose under model uncertainty. The proposed methodology depends on efficient computations, for which a particle swarm optimization (PSO) algorithm can be considered.

Strong Orthogonal Arrays of Strength Two Plus
Boxin Tang, Simon Fraser University

Abstract Strong orthogonal arrays were recently introduced and studied in He and Tang (2013) as a class of space-filling designs for computer experiments. To enjoy the benefits of better space-filling properties, when compared to ordinary orthogonal arrays, strong orthogonal arrays need to have strength three or higher, which may require run sizes that are too large for experimenters to afford. To address this problem, we introduce a new class of arrays, called strong orthogonal arrays of strength two plus. These arrays, while being more economical than strong orthogonal arrays of strength three, still enjoy the better two-dimensional space-filling property of the latter. Among the many results we have obtained on the characterizations and construction of strong orthogonal arrays of strength two plus, worth special mention is their intimate connection with second order saturated designs.

Invited Session 11: Predictive Modeling and Quality Control for Additive Manufacturing

Statistical Quality Monitoring in Metal Additive Manufacturing
Bianca Colosimo, Politechico Di Milano, Italy

Abstract Metal Additive Manufacturing represents a promising solution for digital production of a novel generation of products in many industrial sectors (e.g., aerospace, automotive, biomedical, tooling). This talk describes opportunities and challenges when statistical quality monitoring is applied to metal additive manufacturing. Special focus is devoted to in-situ/in-line solutions.

Statistical Learning Issues in Quality Control for Additive Manufacturing
Qiang Huang, Unversity of Southern California

Abstract Additive manufacturing (AM) enables individualized manufacturing of low-volume products with huge varieties and geometric complexity. Control of 3D shape deformation in AM built products has been a challenging issue, particularly under a cybermanufactuirng environment with diverse fabrication conditions. This talks presents statistical/machine learning issues in quality control for AM and some of our initial results.

Predictive Model Building Across Different Process Conditions and Shapes in 3D Printing
Arman Sabbaghi, Purdue University

Abstract Predictive models for geometric shape deformation constitute an important component in geometric fidelity control for three-dimensional (3D) printing. However, model building is made difficult by the wide variety of possible process conditions and shapes. A methodology that can make full use of data collected on different shapes and conditions, and reduce the haphazard aspect of traditional statistical model building techniques, is necessary in this

context. We develop a new Bayesian procedure based on the effect equivalence and modular deformation features concepts that incorporates all available data for the systematic construction of predictive deformation models. Our method is applied to dramatically facilitate modeling of the multiple deformation profiles that exist in flat cylinders with different types of cavities. Ultimately, our Bayesian approach connects different process conditions and shapes to provide a unified framework for geometric fidelity control in 3D printing.
Statistical Quality Monitoring in Metal Additive Manufacturing

Invited Session 12: Time-series Analysis for Bussiness Problems

Case Sizes and the Bullwhip Effect
Chaitra Nagaraja, Fordham University

Abstract The bullwhip effect is measured by the difference between variability of demand and variability of orders within a supply chain. It has been shown that this effect tends to increase as one moves upstream in the supply chain and this uncertainty causes problems with inventory management. We develop a bullwhip effect for a two-stage supply chain with multivariate, but stationary, demand (e.g., multiple products). In our approach, we require order quantity forecasts to be non-negative integers which are multiples of case size. This stipulation captures the feature that while customers may purchase individual units of a product, orders must be placed by retailers in bundles (i.e., case size, pallet size). A simulation study illustrates the approach.

Statistical Challenges in Forecasting Revenue for Hierarchically Structured Businesses
Julie Novak, IBM T.J. Watson Research Center

Abstract 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.

Dynamic Models for Multivariate Times Series of Counts
Nalini Ravishanker, University of Connecticut

Abstract Discrete-valued time series modeling is emerging as an important research area with diverse applications, as discussed in the recent CRC Handbook of Discrete-valued Time Series. Using Markov Chain Monte Carlo (MCMC) methods for Bayesian hierarchical dynamic modeling of vector time series of counts under a multivariate Poisson sampling distributional assumption may be computationally demanding, especially in high dimensions. An alternate flexible level correlated model (LCM) framework is described in this talk. This enables us to combine different marginal count distributions and to build a hierarchical model for the vector time series of counts, while accounting for association between components of the response vector. We employ the Integrated Nested Laplace Approximation for fast approximate Bayesian modeling using the R-INLA package (r-inla.org). The approach lends itself to application in diverse areas such as ecology, marketing and transportation safety. This talk describes analysis of data from a large multinational pharmaceutical firm on prescription counts for competing drugs in a therapeutic category. Our three-stage analysis enhances computational speed and provides useful guidance to the pharmaceutical firm on their marketing actions.

Invited Session 13: Technometrics Invited Session

Flexible Sequential or Sliced Designs for Computer Experiments
Mingyao Ai, Peking University

Abstract Sequential experiments composed of initial experiments and follow-up experiments are widely adopted for economical computer emulations. Many kinds of Latin hypercubes with good space-filling properties have been constructed for designing the initial computer experiments. However, few works based on Latin hypercubes have focused on the design of the follow-up experiments. Although some constructions of nested Latin hypercube designs can be adapted to sequential designs, the size of the follow-up experiments needs to be a multiple of that of the initial experiments. In this paper, a general method for constructing sequential designs of flexible size is proposed, which allows the combined designs to have good one-dimensional space-filling properties. Moreover, the sampling properties and central limit theorems are derived for these designs. Several improvements of these designs are made to achieve better space-filling properties. The related issues for flexible sliced designs are also addressed. Some simulations are carried out to verify the theoretical results.

Optimization of Multi-Fidelity Computer Experiments via the EQIE Criterion
Xu He, Chinese Academy of Sciences

Abstract Computer experiments based on mathematical models are powerful tools to approximate physical processes. This talk addresses the problem of kriging-based optimization for deterministic computer experiments with

tunable accuracy. Our approach is to use multi-fidelity computer experiments with increasing accuracy levels and a nonstationary Gaussian process model. We propose an optimization scheme that sequentially adds new computer runs by following two criteria. The first criterion, called EQI, scores candidate inputs with given level of accuracy, and the second criterion, called EQIE, scores candidate combinations of inputs and accuracy. From simulation results and a real example using finite element analysis, our method outperforms the expected improvement (EI) criterion which works for single-accuracy experiments.

Real-time Monitoring of High-Dimensional Functional Data Streams via Spatio-Temporal Smooth Sparse Decomposition
Kamran Paynabar, Georgia Institute of Technology

Abstract High dimensional data monitoring and diagnosis has recently attracted increasing attentions. However, existing process monitoring methods fail to fully utilize the information of high dimensional data streams due to their complex characteristics including the large dimensionality, spatio-temporal correlation structure, and non-stationarity. In this paper, we propose a novel process monitoring methodology for high-dimensional data streams including profiles and images that can effectively address foregoing challenges. We introduce spatio-temporal smooth sparse decomposition (ST-SSD), which serves as a dimension reduction and denoting technique by decomposing the original tensor into the functional mean, sparse anomalies, and random noises. ST-SSD is followed by a sequential likelihood ratio test on extracted anomalies for process monitoring. To enable real-time implementation of the proposed methodology, recursive estimation procedures for ST-SSD are developed. ST-SSD also provides useful diagnostics information about the location of change in the functional mean. The proposed methodology is validated through various simulations and real case studies.

Invited Session 14: Reliability Modeling and Applications

Parallel Computing and Network Analytics for Fast Industrial Internet-of-Things (IIOT) Machine Information Processing and Condition Monitoring
Chen Kan, Penn State University

Abstract In the past decade, rapid advancements of sensing and communication technology bring the new wave of Industrial Internet of Things (IIOT). However, realizing the full potential of IIOT depends on the development of new methodologies for big data analytics. Existing approaches are limited in their ability to effectively extract pertinent knowledge about manufacturing operations from the large volume of IIOT data of networked machines. This paper presents new parallel algorithms for large-scale IIOT machine

information processing, network modeling, condition monitoring, and fault diagnosis. Experimental results show that the new parallelized algorithm efficiently and effectively characterizes the variations of machine signatures for network modeling and monitoring. This new approach shows strong potentials for optimal machine scheduling and maintenance in the large-scale IIOT context.

Analysis of Reliability Experiments with Blocking
Xiao Liu, University of Arkansas/IBM Thomas J. Watson Research Center

Abstract Many reliability life tests contain blocking or subsampling. This is often the case when treatments are directly applied to test stands rather than individual specimen, or, when specimens come from different production batches. Incorrectly assuming completely randomized design underestimates the standard errors due to overstating the true experimental degrees of freedom. In this talk, a survey of existing approaches to analyzing reliability life tests data under subsampling is conducted. We propose a method that integrates the idea of frailty, which accounts for the subsampling effect, and the technique of multiple imputation for analyzing experimental data. A step-by-step description of the approach is presented, followed by a numerical example based on a popular reliability dataset. Finally, comprehensive comparison studies between the proposed method and existing methods are conducted

Learning from Multi-Modality Multi-Resolution Time Series Data: an Optimization Approach
Yada Zhu, IBM Research

Abstract Many complex real applications involve the collection of time series data with multiple modalities and of multiple resolutions. For example, in aluminum smelting processes, the recorded process variables typically reflect various aspects of these processes, such as pressure and temperature, and they are often obtained with different time resolutions, such as every 5 minutes and every day. How can we effectively leverage both the multi-modality property and the multi-resolution property of the data for the sake of more accurate prediction of key process indicators (e.g., the cell temperature of the aluminum smelting processes)?
Different from existing techniques, which can only model the multi-modality property or the multi-resolution property, in this paper, for the first time, we propose to jointly model the two properties such that the prediction results are consistent across multiple modalities and multiple resolutions. To this end, we construct an optimization frame- work, which is based on a novel regularizer imposing such consistency. Then, we design an effective and efficient optimization algorithm based on randomized block coordinate descent. Its performance is evaluated on both synthetic and real data sets, outperforming state-of-the-art techniques.

Contributed Session 1: New Strategies for Collection and Analysis of Data to Solve Modern-Day Scientific and Engineering Problems

Compromise Designs Under Baseline Parametrization
Ruwan Chamara Karunanayaka, Simon Fraser University

Abstract We consider estimation of main effects using two-level fractional factorial designs under the baseline parametrization. Previous work in the area indicates that orthogonal arrays are more efficient than one-factor-at-a-time designs whereas the latter are better than the former in terms of minimizing the bias due to non-negligible interactions. Using efficiency criteria, this paper examines a class of compromise designs obtained by adding runs to one-factor-at-a-time designs. A theoretical result is established for the case of adding one run. For adding two or more runs, we develop a complete search algorithm to find optimal compromise designs.

Using Particle Swarm Optimization to Search for Optimal Designs for Mixed Factor Experiments with Binary Response
Abhyuday Mandal, University of Georgia

Abstract Identifying optimal designs for generalized linear models with a binary response can be a challenging task, especially when there are both continuous and discrete independent factors in the model. Theoretical results rarely exist for such models, and for the handful that do, they come with restrictive assumptions. Here we illustrate the use of particle swarm optimization (PSO) to search for locally D-optimal designs for generalized linear models with discrete and continuous factors and a binary outcome and demonstrates that PSO can be an effective method. We provide two real applications using PSO to identify designs for experiments with mixed factors: one to redesign an odor removal study and the second to find an optimal design for an electrostatic discharge study. In both cases we show that the D-efficiencies of the designs found by PSO are much higher than the implemented designs. In addition, we show PSO can efficiently find D-optimal designs on a prototype or an irregularly shaped design space, provide insights on the existence of minimally supported optimal designs, and evaluate sensitivity of the D-optimal design to mis-specifications in the link function. We also study the robustness of locally optimal designs with respect to the mis-specification of initial guesses of the model parameter values, and also identify pseudo-Bayesian designs. (joint research with Joshua Lukemire and Weng Kee Wong).

Sequential Sampling Enhanced Composite Likelihood Approach to Estimation of Social Intercorrelations in Large-Scale Networks
Youran Qi, University of Wisconsin

Abstract The increasing access to large social network data has generated substantial interest in the marketing community. However, due to its large scale, traditional analysis methods often become inadequate. In this paper, we propose a sequential sampling enhanced composite likelihood approach for efficient estimation of social intercorrelations in large-scale networks using the spatial model. The proposed approach sequentially takes small samples from the network, and adaptively improves model parameter estimates through learnings obtained from previous samples. In comparison to population-based maximum likelihood estimation that is computationally prohibitive when the network size is large, the proposed approach makes it computationally feasible to analyze large networks and provide efficient estimation of social intercorrelations among members in large networks. In comparison to sample-based estimation that relies on information purely from the sample and produces underestimation bias in social intercorrelation estimates, the proposed approach effectively uses information from the population without compromising computation efficiency. Through simulation studies based on simulated networks and real networks, we demonstrate significant advantages of the proposed approach over benchmark estimation methods and discuss managerial implications.

Semi-parametric Adjustment to Computer Models
Rui Tuo, Georgia Institute of Technology

Abstract Computer simulations are widely used in scientific exploration and engineering designs. However, computer outputs usually do not match the reality perfectly because the computer models are built under certain simplifications and approximations. When physical observations are also available, statistical methods can be applied to estimate the discrepancy between the computer output and the physical response. In this article, we propose a semi-parametric method for statistical adjustments to computer models. We use three numerical studies and a real example to examine the predictive performance of the proposed method. The results show that the proposed method outperforms existing methods.

Invited Session 15: Design of experiments II

Analyzing Definitive Screening Designs Taking Advantage of their Structure
Bradley Jones, JMP

Abstract Designed experiments often have strong symmetry (such as orthogonal columns). This suggests that analytical methods for designed experiments could profitably take advantage of what is already known about their structure. One might call this idea design oriented modeling. Definitive Screening Designs (DSDs) have a special structure with many desirable

properties. They have orthogonal main effects, and main effects are also orthogonal to all second order effects. DSDs with more than five factors project onto any three factors to enable efficient fitting of a full quadratic model. However, analytical methods for DSDs employ generic tools invented for regression analysis of observational data. These approaches do not take advantage of all the useful structure that DSDs provide. This talk introduces an analytical approach for DSDs that does take explicit advantage of the special structure of DSDs. To make the methodology clear, the presentation will provide a step by step procedure for analysis using specific examples.

The Construction of Space-Filling Designs with Good Uniformities in Multiple Dimensions via Factor Subgroup Collapse
Frederick Phoa, Academia Sinica, Taiwan

Abstract This work introduces a new class of space-filling designs optimized under a new multi-dimensional space-filling property called {\it geometric strength}. We propose a systematic construction method via techniques in Galois field for this new class of designs. In specific, the factor levels in a regular design are collapsed and the strength of the collapsed design is enhanced. The reversed process to relabel factor levels of the regular design improves its space-filling property. This method is more efficient than the existing methods via level permutations, especially when the number of factor levels is large. When two collapsers are indistinguishable in terms of the strength of the collapsed designs, we propose a new criterion called maximal strength efficiency. It not only maximizes the strength of the collapsed design, but also maximizes the proportion of the projected sub-designs that are full factorials.

Closed-Loop Automatic Experimentation for Bayesian Optimisation
David Woods, University of Southampton

Abstract Automated experimental systems, involving minimal human intervention, are becoming more popular and common, providing economical and fast data collection. We discuss some statistical issues around the design of experiments and data modelling for such systems. Our application is to “closed-loop” optimisation of chemical processes, where automation of reaction synthesis, chemical analysis and statistical design and modelling increases lab efficiency and allows 24/7 use of equipment.
Our approach uses nonparametric regression modelling, specifically Gaussian process regression, to allow flexible and robust modelling of potentially complex relationships between reaction conditions and measured responses. A Bayesian approach is adopted to uncertainty quantification, facilitated through computationally efficient Sequential Monte Carlo algorithms for the approximation of the posterior predictive distribution. We propose a new criterion, Expected Gain in Utility (EGU), for optimisation of a noisy response via fully-sequential design of experiments, and we compare the performance of

EGU to extensions of the Expected Improvement criterion, which is popular for optimisation of deterministic functions. We also show how the modelling and design can be adapted to identify, and then down-weight, potentially outlying observations to obtain a more robust analysis.
Joint work with Tim Waite (University of Manchester, UK)

Invited Session 16: Recent Advances in Statistical Learning

Social Networks of Statisticians
Zheng Ke, The University of Chicago

Abstract We have collected a data set for the social networks of statisticians. The data set consists of the meta information of about 70,000 papers in representative statistics journals. Our data collection project (Phase II) is a continuation of the recent data collection project by Ji and Jin (Phase I). We investigate the Phase II data and report some Exploratory Data Analysis (EDA) results. In particular, we discuss the overall productivity, journal-journal citation exchanges, and citation patterns of individual papers.
The data sets also pose many problems, to solve which we need more sophisticated methods. One such problem is mixed membership estimation. We propose Mixed-SCORE as a new spectral method for mixed membership estimation. At the heart of Mixed-SCORE is a matrix of entry-wise ratios, formed by dividing the first few eigenvectors of the network adjacency matrix over the leading eigenvector of the same matrix in an entry-wise fashion. The main surprise is that the rows of the entry-wise ratio matrix form a cloud of points in a low-dimensional space with the silhouette of a simplex, which simplex carries all information we need for estimating the memberships. We apply Mixed SCORE to several networks constructed with the Phase I data and obtain meaningful results. We propose a Degree-Corrected Mixed Membership model and use it to solidify our discoveries theoretically.
This is joint work with Pengsheng Ji (U George), Jiashun Jin (CMU) and Shengming Luo (CMU).

Hypothesis Testing for Stochastic Block Models via Linear Spectral Statistics
Zongming Ma, University of Pennsylvania

Abstract In this talk, we present CLTs for linear spectral statistics under stochastic block models, which leads to useful test statistics for distinguishing Erdos-Renyi random graphs and those from stochastic block models. The resulting tests work whenever the average node degree goes to infinity as the graph size grows, and they have power in both the contiguous and the singular regimes. These tests can potentially be used for determining the number of communities in stochastic block models.
The talk is based on joint work with Debapratim Banerjee.

Recent Developments in High-Dimensional Tensor Regression Problems

Garvesh Raskutti, University of Wisconsin

Abstract: Tensor data is arising more and more frequently in many scientific applications including neuro-imaging, text data, and many others. However the theory and methodology involving tensor data beyond simple vectorization schemes is generally under-developed. In this talk I present different approaches for solving high-dimensional tensor regression problems under low-dimensional structural assumptions: (1) a convex regularization scheme and (2) a non-convex but computable projection-based scheme where the underlying tensor lies in an unknown low-dimensional subspace. In particular, I derive general risk bounds of the resulting estimate under fairly general dependence structure among covariates. These general bounds provide useful upper bounds on rates of convergence for a number of fundamental statistical models of interest including multi-response regression, vector auto-regressive models, low-rank tensor models and pairwise interaction models. Moreover, in many of these settings we prove that the resulting estimates are minimax optimal. I finish by discussing some open computational and statistical issues associated with low-rank tensor regression.

Stochastic Methods for Composite Optimization Problems
Feng Ruan, Stanford University

Abstract We consider minimization of stochastic functionals that are compositions of a (potentially) non-smooth convex function h and smooth function c. We develop two stochastic methods---a stochastic prox-linear algorithm and a stochastic (generalized) sub-gradient procedure---and prove that, under mild technical conditions, each converges to first-order stationary points of the stochastic objective. We provide experiments further investigating our methods on non-smooth phase retrieval problems, the experiments indicate the practical effectiveness of the procedures.

Invited Session 17: Machine Learning in Complex Systems

Comparison of Gaussian process modeling software
Collin Erickson, Northwestern University

Abstract Gaussian process fitting, or kriging, is often used to create a model from a set of data. Many available software packages do this, but we show that very different results can be obtained from different packages even when using the same data and model. Eight different fitting packages that run on four different platforms are compared using various data functions and data sets that reveal there are stark differences between the packages. In addition to comparing the prediction accuracy, the predictive variance—which is important for evaluating precision of predictions and is often used in stopping criteria—is also evaluated.

On Reject and Refine Options in Multicategory Classification
Xingye Qiao, Binghamton University

Abstract In many real applications of statistical learning, a decision made from misclassification can be too costly to afford; in this case, a reject option, which defers the decision until further investigation is conducted, is often preferred. In recent years, there has been much development for binary classification with a reject option. Yet, little progress has been made for the multicategory case. In this article, we propose margin-based multicategory classification methods with a reject option. In addition, and more importantly, we introduce a new and unique refine option for the multicategory problem, where the class of an observation is predicted to be from a set of class labels, whose cardinality is not necessarily one. The main advantage of both options lies in their capacity of identifying error-prone observations. Moreover, the refine option can provide more constructive information for classification by effectively ruling out implausible classes. Efficient implementations have been developed for the proposed methods. On the theoretical side, we offer a novel statistical learning theory and show a fast convergence rate of the excess ℓ-risk of our methods with emphasis on diverging dimensionality and number of classes. The results can be further improved under a low noise assumption and be generalized to the excess 0-d-1 risk. Finite-sample upper bounds for the reject and reject/refine rates are also provided.

Physical-Statistical Modeling and Regularization of High-Dimensional Dynamic Systems
Bing Yao, Penn State University

Abstract This talk presents a novel physics-driven spatiotemporal regularization (STRE) method for high-dimensional predictive modeling in complex space-time systems. First, we developed realistic models of torso-heart geometry, and utilized the boundary element method and physics-based principles to derive the parameter matrix R, which captures the physics-based interrelationship between time-varying explanatory and response variables that are distributed in the space. Second, we developed a physical-statistical approach that integrates physics-derived parameter matrix R with a spatiotemporal regularization method to build the high-dimensional predictive model. Third, we designed a new method of dipole multiplicative update, inspired by the dipole assumption in electro-dynamic physics, to solve the generalized spatiotemporal regularization problems. The STRE model is implemented to predict the time-varying distribution of electric potentials on the heart surface based on the electrocardiogram (ECG) data from the body sensor network. The model performance is evaluated and validated in both a simulated two-sphere geometry and a realistic torso-heart geometry. Experimental results show that the STRE model significantly outperforms other regularization models that are widely used in current practice such as

Tikhonov zero-order, Tikhonov first-order and L1 first-order regularization methods.

A Nonparametric Approach for Partial Areas under ROC Curves
Yichuan Zhao, Georgia State University

Abstract The receiver operating characteristic (ROC) curve is a well-known measure of the performance of a classification method. Interest may only pertain to a specific region of the curve and, in this case, the partial area under the ROC curve (pAUC) provides a useful summary measure. Related measures such as the ordinal dominance curve (ODC) and the partial area under the ODC (pODC) are frequently of interest as well. Based on a novel estimator of pAUC proposed by Wang and Chang (2011), we develop nonparametric approaches to the pAUC and pODC using normal approximation, the jackknife and the jackknife empirical likelihood. A simulation study demonstrates the flaws of the existing method and shows proposed methods perform well. Simulations also substantiate the consistency of our jackknife variance estimator. The Pancreatic Cancer Serum Biomarker data set is used to illustrate the proposed methods.

Banquet Speech

Big Data, Big Surprises?
Xiao-Li Meng, Department of Statistics, Harvard University

Abstract The phrase “Big Data” has greatly raised expectations of what we can learn about ourselves and the world in which we live or will live. It appears to have also boosted general trust in empirical findings, because it seems to be common sense that the more data, the more reliable are our results. Unfortunately, this common sense conception can be falsified mathematically even for methods such as the time-honored ordinary least squares regressions, and the issue does not go away even when one has infinite amount of data (Meng and Xie, 2014). Furthermore, whereas the size of data is a common indicator of the amount of information, what matters far more is the quality of data. A largely overlooked statistical identity, a potential candidate for the statistical counterpart to the beautiful Euler identity, reveals that trading quantity for quality in statistical estimation is a mathematically demonstrable doomed game (Meng, 2017). Without taking into account the data quality, Big Data can do more harm than good because of the drastically inflated precision assessment, and hence the gross overconfidence, which minimally can give us serious surprises when the reality unfolds, as illustrated by the 2016 US election.

Invited Session 18: High Dimensional and Complex Time Series

Regularized Estimation and Testing for High-Dimensional Multi-Block Vector Autorgressive Models
George Michailidis, University of Florida

Abstract Dynamical systems comprising of multiple components originate in many scientific areas. A pertinent example is the interactions between financial assets and macroeconomic indicators, which has been studied at an aggregate level in the macroeconomics literature. In this talk, we consider a multi-block linear dynamic system with Granger-causal ordering between blocks, wherein the blocks temporal dynamics are described by vector autoregressive processes and are influenced by blocks higher in the system hierarchy. We obtain the maximum likelihood estimator for the posited model for Gaussian data in the high-dimensional setting with appropriate regularization schemes. To optimize the non-convex likelihood function, we develop an iterative algorithm with convergence guarantees. We establish theoretical properties of the maximum likelihood estimates, leveraging the decomposability of the regularizers and a careful analysis of the iterates of the proposed algorithm. Finally, we develop testing procedures for the null hypothesis of whether a block ``Granger-causes" another block of variables. The performance of the model and the testing procedures are evaluated on synthetic data, and illustrated on a data set involving log-returns of the US S&P100 component stocks and key macroeconomic variables for the 2001--16 period.

Robust Testing and Variable Selection in High-Dimensional Time Series
Ruey Tsay, University of Chicago

Abstract This talk is concerned with analysis of high-dimensional time series. We consider two main issues in analysis of high-dimensional time series. The first issue is testing for serial correlations when the dimension is high and the second issue is variable selection. Limiting distributions of the statistics used in testing and variable selection are derived. Some finite sample corrections are also introduced. We use simulations to demonstrate the performance of the proposed statistics and to compare them with other existing methods, including LASSO. Real examples are also used.

Unsupervised Self-Normalized Change-Point Testing for Time Series
Ting Zhang, Boston University

Abstract We propose a new self-normalized method for testing change points in the time series setting. Self-normalization has been celebrated for its ability to avoid direct estimation of the nuisance asymptotic variance and its flexibility

of being generalized to handle quantities other than the mean. However, it was developed and mainly studied for constructing confidence intervals for quantities associated with a stationary time series, and its adaptation to change-point testing can be nontrivial as direct implementation can lead to tests with nonmonotonic power. Compared with existing results on using self-normalization in this direction, the current paper proposes a new self-normalized change-point test that does not require prespecifying the number of total change points and is thus unsupervised. In addition, we propose a new contrast-based approach in generalizing self-normalized statistics to handle quantities other than the mean, which is specifically tailored for change-point testing. Monte Carlo simulations are presented to illustrate the finite-sample performance of the proposed method.

Contributed Session 2: Recent Advances in Design and Analysis of Classical Experiments

Cmenet - a New Method for Bi-Level Variable Selection of Conditional Main Effects
Simon Mak, Georgia Institute of Technology

This paper presents a novel method for selecting main effects and a set of reparametrized predictors called conditional main effects (CMEs), which capture the conditional effect of a factor at a fixed level of another factor. CMEs represent highly interpretable phenomena for a wide range of applications in engineering, social sciences and genomics. The challenge in model selection lies in the grouped collinearity structure of CMEs, which can cause poor selection and prediction performance for existing methods. We propose a new method called cmenet, which employs coordinate descent and two principles called CME coupling and reduction to efficiently perform model selection. Simulation studies demonstrate the improved performance of cmenet over existing selection methods, such as the LASSO and SparseNet. Applied to a gene association study on fly wing shape, cmenet not only provides improved predictive performance over existing techniques, but also reveals important insight on gene activation behavior. Efficient implementations of our algorithms are available in the R package cmenet in CRAN.

D-Optimal Mixture Designs for Ordinal Responses
Michelle Mancenido, Arizona State University

Mixture experiments are prevalent in industrial and scientific applications where the settings of the experimental factors are dependent on each other and a totality constraint is imposed on the factors. Examples include experiments involving mixtures of chemicals, gases, or drug components. This talk focuses on the design of D-optimal mixture experiments for ordinal responses. Ordinal responses often arise when the experimental outcome cannot be measured in the numeric and continuous scale, such as in patients' assessment of pain in a

clinical trial or the intensity of fragrance of a new perfume in consumer perception studies. We discuss the computational issues in constructing D-optimal mixture designs and propose an adapted exchange heuristic for calculating local and robust designs for ordinal responses. Finally, standard mixture designs are evaluated vis-a- vis the D-optimal designs and insights are provided on the use of standard designs as surrogates.

Supersaturated Designs Robust to Two-Factor Interactions
Chenlu Shi, Simon Fraser University

This paper proposes and studies a new class of supersaturated designs, foldover supersaturated designs. Such designs allow us to screen out active main effects without making the assumption that interactions are negligible. A lower bound for E(s^2) values of foldover supersaturated designs is derived and a comparison is made between the existing supersaturated designs and our designs with respect to estimation of main effects under the presence of interactions. We then present a method using regular designs for constructing foldover supersaturated designs that achieve the lower bound for E(s^2). Further we develop a procedure of searching good E(s^2)-optimal foldover supersaturated designs from good Hadamard matrices in terms of minimizing the pairwise correlations.

A Method for Identifying Dominant Effects in Designs with Complex Aliasing
Fasheng Sun, Northeast Normal University

For saving on costs, we usually use a design with a small run size, for example Plackett-Burman (PB) design, to conduct an experiment. But such a designs has complex aliasing pattern, and it is very difficult to identify the dominant effects. This paper suggests an analysis method for such designs, based on the decomposition of the variance of response. This method is easy to perform and also works when the number of the effects need to be identified is more than the run size. Thus the method can also be used for screening the active effects in supersaturated designs.

Contributed Session 3: New Statistical and Machine Learning Methodologies Motivated by Cutting Edge Scientific Applications

Anisotropic Functional Laplace De-Convolution and Its Application to Dynamic Contrast Enhanced Imaging
Rida Benhaddou, Ohio University

In the present paper we consider the problem of estimating a three-dimensional function *f* based on observations from its noisy convolution. We construct an adaptive wavelet-Laguerre estimator of *f*, derive minimax lower bounds for the L2-risk when *f* belongs to three-dimensional Laguerre-Sobolev ball and

demonstrate that the wavelet-Laguerre estimator is adaptive and asymptotically near-optimal within a logarithmic factor in a wide range of Laguerre-Sobolev balls. We carry out a limited simulations study and show that the estimator performs well in a finite simulation setting. Finally, we present a real data example which motivates the problem studied in the paper.

The Blessing of Derivatives in Nonparametric Estimation
Xiaowu Dai, University of Wisconsin Madison

We study smoothing regularization methods for incorporating derivatives in nonparametric function estimation. Data with derivative info arise in economics, engineering, uncertainty quantification and many other fields. We obtained new results to show that the dimension of the estimation of a multidimensional function can be reduced if the first-order partial derivatives of the function are available. Also, we established that the regularization with incorporation of derivative data is rate-optimal and adapts to unknown smoothness up to an order related to the given kernel. We provide theoretical results to show that in finite sample cases, the proposed regularization estimator produces smaller mean squared error than least squares estimators used in previous studies. The proposed estimation procedure is easy to implement and generally applicable to a wide range of kernels. Numerical examples are provided to corroborate the derived theoretical results.

Bayesian Analysis of Traffic Flow Data
Vadim O. Sokolov, George Mason University

Many business and engineering problems ranging from identifying cyber attacks to real-time financial analytics require analysis of non-linear and non-Gaussian data, in order to infer the parameters of underlying data generating process. In this paper we develop a Bayesian algorithm for nonlinear/non-Gaussian filtering and learning, and apply it to analysis of traffic flow data. We develop a particle filtering and learning algorithm to estimate the state of current traffic density and the parameters of the non-liner traffic flow model. These inputs are related to the so-called fundamental diagram, which describes the relationship between traffic flow and density. Our methodology allows for real-time updating of the posterior uncertainty for the critical density and capacity parameters. We provide a real-time data analysis of how to learn the drop in capacity, as a result of a major traffic accident. Our algorithm allows us to accurately assess the uncertainty of the current traffic state at shock waves, where the uncertainty is a mixture distribution. We show that Bayesian learning can correct the estimation bias that is present in the model with fixed parameters.

Generalized Support Vector Data Description with Bayesian Framework
Mehmet Turkoz, Rutgers University

Many anomaly detection procedures assume that the target data consist of only one class. However, in many real life problems, the target data may have more than one class. Therefore, existing anomaly detection procedures such as SVDD may fail to describe the target data properly. In addition, the results of the SVDD are deterministic, which is not desirable compared to having probabilistic interpretation. To overcome these limitations of traditional SVDD, in this research, Bayesian n-SVDD is introduced. We also evaluate and demonstrate the performance of the proposed approach using data sets with multi-class

Invited Session 19: Emerging Statistical Inference Methods in the Era of Data Science

On Integrative Learning of Mixed and Incomplete Data
Kun Chen, University of Connecticut

Abstract Multivariate outcomes together with multivariate features of possibly high dimensionality have been routinely produced from various fields. In many problems, the collected outcomes are of mixed types, including continuous measurements, binary indicators and counts, and the data may also subject to substantial missing values. Regardless of their types, these mixed outcomes are often interrelated, representing diverse views of the same underlying data generation mechanism. As such, an integrative multivariate learning can be beneficial. We develop a mixed-outcome reduced rank regression, which integrates mixed and incomplete outcomes belonging to the exponential-dispersion family by assuming that all the outcomes are associated through a shared low-dimensional subspace spanned by the high-dimensional features. A general regularized estimation criterion is proposed, and a unified algorithm with convergence guarantee is developed for optimization. We establish non-asymptotic performance bound for the proposed estimators under a general sampling scheme of missing. The effectiveness of our approach is demonstrated by simulation studies and an application on predicting health-related outcomes in longitudinal studies of aging. Extensions and other strategies for modeling mixed-type data including sequential sparse feature extraction and mixture models will be discussed.

Recent Development on CPO Statistics in Joint Modeling of Longitudinal and Survival Data
Ming-Hui Chen, University of Connecticut

Abstract Joint models for longitudinal and survival data are routinely used in clinical trials or other studies to assess a treatment effect while accounting for longitudinal measures such as patient-reported outcomes (PROs). In the Bayesian framework, the logarithm of the pseudo marginal likelihood (LPML) is a well-known Bayesian criterion for comparing joint models. However, this

criterion does not provide separate assessments of each component of the joint model. To circumvent this issue, we develop a novel decomposition of the Conditional Predictive Ordinate (CPO) statistics and consequently LPML to assess the fit of the longitudinal and survival components of the joint model, separately. Based on this decomposition, we then propose a Bayesian model assessment criterion, namely, ΔLPML, to determine the importance and contribution of the longitudinal (survival) data to the model fit of the survival (longitudinal) data. Moreover, we develop an efficient Monte Carlo method for computing CPO statistics in the joint modeling setting. The proposed methodology is applied to a case study in mesothelioma. This is a part of the joint work with Danjie Zhang, Joseph G. Ibrahim, Mark E. Boye, and Wei Shen.

Testing for Homogeneity in Mixture Models

Yong Chen, University of Pennsylvania

Abstract DNA methylation plays an important role in the development of many types of cancer. Identifying differentially methylated loci between cancer and normal patients is one of the central tasks to understand the impacts of methylation process on cancer development. In this talk, we motivate a class of new mixture models from the UK ovarian cancer population study. The proposed models are devised to account for several key features of methylation data in cancer research, including the heterogeneity in methylation distributions among cancer patients. We propose two statistical tests to detect differentially methylated loci, which are showed to achieve a much-improved statistical power and/or model robustness compared to existing methods. The statistical techniques used to handle irregularity of the hypothesis-testing problem in this mixture model context will also be described.

Analysis Methods in the Presence of Mixed Measurement Error and Misclassification in Covariates

Grace Yi, University of Waterloo

Abstract Covariate measurement imprecision or errors arise frequently in many areas. It is well known that ignoring such errors can substantially degrade the quality of inference or even yield erroneous results. Although in practice both covariates subject to measurement error and covariates subject to misclassification can occur, research attention in the literature has mainly focused on addressing either one of these problems separately. In this paper, we develop estimation and inference methods that accommodate both characteristics simultaneously. Specifically, we consider measurement error and misclassification in generalized linear models under the scenario that an external validation study is available and develop several functional and structural methods. This is joint work with Yanyuan Ma, Donna Spiegelman and Raymond J. Carroll.

Contributed Session 4: New Developments in Space-filling Designs

Optimal Space-filling Latin Hypercube Designs Based on Good Lattice Point Sets
Lin Wang, UCLA

Space-filling Latin hypercube designs are commonly used for computer experiments. We construct three classes of space-filling Latin hypercube designs via non-linear transformations of good lattice point sets. One class of these designs are optimal under the maximin distance criterion, while two other classes are asymptotically optimal under this criterion. Moreover, these designs are also shown to have low pairwise correlations between columns.

Space-filling Properties of Mirror-symmetric Orthogonal Designs
Yaping Wang, Peking University

U-type designs, especially Latin hypercube designs, have been extensively used in computer experiments. Column-orthogonality and maximin distance are two popular criteria for optimizing such designs. Compared with many theories of column orthogonal designs, less results have been obtained, due to the complexity, for exact maximin designs except for the two-dimensional cases. In this paper, we theoretically characterize a broad class of maximin distance designs by giving new bounds on the minimum inter-site distance for mirror-symmetric and general U-type designs, respectively. In particular, it is shown that the two criteria, maximin distance and columnorthogonality, are closely related and coincide within a variety of parameters. Some exact and nearly maximin designs are constructed and listed for practical use.

Construction of Maximin Distance Latin Squares and Related Latin Hypercube Designs
Qian Xiao, UCLA

Maximin distance Latin hypercube designs are widely used in computer experiments, yet their construction is challenging. Based on number theory and finite fields, we propose three algebraic methods to construct maximin distance Latin squares, as special Latin hypercube designs. We develop lower bounds on their minimum distances. The resulting Latin squares and related Latin hypercube designs have larger minimum distances than existing ones, and are especially appealing for high dimensional applications.

Construction of Nearly Uniform Designs on Irregular Regions
Jianfeng Yang, Nankai University

The uniform design is a kind of important experimental design which has great practical value in production and living. Most existing literatures on this topic focus on the construction of uniform designs on regular regions. However, because of the complexity of practical situations, the irregular design region is more common in real life. In this paper, an algorithm is proposed to the

construction of nearly uniform designs on irregular regions. The basic idea is to make use of uniform designs on a larger regular region with the irregular region being a subregion. Some theoretical justifications on the proposed algorithm are provided. Both the comparisons with the existing results and a real-life example show that our proposed algorithm is effective.

Contributed Session 5: New Paradigms and Approaches in Modern-Day Process Monitoring

A Novel Pattern-Frequency Tree Approach for Transition Analysis and Anomaly Detection in Nonlinear and Nonstationary Systems
Cheng-Bang Chen, Penn State University

Identifying the irregular transition regions of the normal process signals is an important task for system monitoring and control. The failure to identify anomalous patterns in dynamic systems can lead to the occurrence of catastrophic events and result in higher cost. The prior research attempted to use multiple sensors for a closer monitoring of the system dynamics. However, realizing full utilization of multiple sensors without the normality assumptions and dimensionality reduction remains a research challenge to build control schemes. This paper aims to fill the gap and develops a novel approach utilizing the information from pattern-frequency trees to detect the abnormal regions of a high dimensional complex system. First, we use state space method to represent the quasi-periodic signals of a dynamic complex system. Then, we propose a spatial indexing method, Hyperoctree State space Aggregation Segmentation (HSAS), to delineate the high-dimensional dynamic processes in a continuous state space. With the Piecewise Aggregate Approximation (PAA), we extract the patterns from the transitory signals and then construct a pattern-frequency tree. Finally, we leverage the pattern-frequency distribution information to develop a k-Maximin deviation algorithm for effective and efficient detection of process anomalies. Experimental results demonstrate that the proposed method performs better than the conventional methods in multi-sensor settings in high-dimensional environments.

Sensor Fusion and On-line Monitoring of Friction Stir Blind Riveting for Lightweight Materials Manufacturing
Zhe Gao, Rutgers University

Friction stir blind riveting (FSBR) is a recently developed manufacturing process for joining automotive lightweight materials. During FSBR, a blind rivet rotates at high speed, contacts with the upper sheet or workpiece of a lap joint, and then penetrates the workpieces. Using FSBR to join carbon fiber-reinforced polymer (CFRP) composite and aluminum alloy sheets has been studied experimentally, however, the quantitative relationship between FSBR and joint quality/strength remains unclear. To gain a better understanding of FSBR lightweight materials manufacturing, the proposed method effectively models

this relationship by integrating data de-noising, dimension reduction, feature extraction, feature selection, and classifier fusion. Engineering-based features are extracted directly from the FSBR penetration force and torque signals; data-driven features are extracted from multiple heterogeneous sensor signals using lower-rank tensor decomposition (LRTD) algorithms. Multiple LRTD algorithms are implemented and their effectiveness is compared using various datasets. Finally, regression models and kernel support vector machines (SVMs) are trained and fused for online quality prediction. The proposed method is demonstrated with both simulated and real data from FSBR.
Keywords: Online monitoring, Sensor fusion, Tensor decomposition, Feature selection, Quality prediction, Lightweight manufacturing

Nonparametric Change-Point Detection for Process Monitoring and Prognostics in Advanced Manufacturing
Shenghan Guo, Rutgers University

It is essential to detect significant changes to the production process. Most manufacturing data are complicated and cannot be modeled by any widely-used probability distributions. In this study, we develop a monitoring and prognostics method that implements non-parametric change-point detection methods on non-stationary time series data collected from a real automobile manufacturing system. For offline change-point detection, we integrate three non-parametric methods: the least absolute shrinkage and selection operator (LASSO), the threshold least absolute shrinkage and selection operator (TLASSO), and the wild binary segmentation (WBS); for online change-point detection, we implement the modified sequential change-point detection. Comparing to LASSO, TLASSO gives fewer change points, as only abrupt changes will be considered in this method. Although both LASSO and TLASSO show good performance in change-point detection, they suffer from a low computational efficiency, which is a relative advantage of WBS. The modified sequential change-point detection method is effective for online monitoring, given that a moderate amount of delay is acceptable. These nonparametric change-point detection methods are compared in case studies with real manufacturing data. Detected change points are then used for online process monitoring and prognostics. Expected benefits of the proposed method, such as reducing unnecessary shut-downs by 20%, are assessed in simulation.

Statistical Process Monitoring of High-Dimensional Processes via Ridge
Sangahn Kim, Rutgers University

As the number of quality characteristics to be monitored increases in those complex processes, the simultaneous monitoring becomes less sensitive to the out-of-control signals especially when only a few variables are responsible for abnormal situation. We introduce a new process control chart for monitoring high-dimensional processes based on the ridge penalizing likelihood. The accurate probability distributions under null and alternative hypotheses are

obtained. In addition, we find out several theoretical properties of the proposed method, and finally demonstrate the proposed chart performs well in monitoring high-dimensional processes.