AISC – 2018
International Conference
on
Advances in Interdisciplinary Statistics *and* Combinatorics

October 5 - 7, 2018

Conference Program and Abstracts

ASA

Promoting the Practice and Profession of Statistics™

UNC GREENSBORO

AISC–2018 International Conference on Advances in Interdisciplinary Statistics and Combinatorics

A Conference Sponsored by
The North Carolina Chapter of the American Statistical Association

October 5-7, 2018

The University of North Carolina at Greensboro

Contents

Conference Sponsors

Institute for Mathematics and Its Applications

The University of North Carolina at Greensboro

The North Carolina Chapter of the American Statistical Association

The American Statistical Association

Pearson

RHO

SAS

National Institute of Statistical Sciences

Springer

International Conference on Advances in Interdisciplinary Statistics and Combinatorics AISC – 2018

Local Organizing Committee

Ratnasingham Shivaji (Advisor)
Head and Helen Barton Excellence Professor
Department of Mathematics and Statistics, UNCG

Sat Gupta (Organizer and Chair)
Department of Mathematics and Statistics, UNCG

Haimeng Zhang (Co-Organizer)
Department of Mathematics and Statistics, UNCG

Xiaoli Gao
Department of Mathematics and Statistics, UNCG

Scott Richter
Department of Mathematics and Statistics, UNCG

Igor Erovenko
Department of Mathematics and Statistics, UNCG

Somya Mohanty
Department of Computer Science, UNCG

Scientific Program Committee

Sat Gupta (Chair)
Department of Mathematics and Statistics
UNCG
sngupta@uncg.edu

Hrishikesh Chakraborty
Associate Director, Clinical Trial Statistics
Duke Clinical Research Institute
Duke University
rishi.c@duke.edu

Angela Dean
Department of Statistics
Ohio State University
amd@stat.ohio-state.edu

Xiaoli Gao
Department of Mathematics and Statistics
UNCG
x_gao2@uncg.edu

Sujit Ghosh
Department of Statistics
NC State University

Benjamin Kedem
Department of Mathematics
University of Maryland
bnk@math.umd.edu

Prashanti Manda
Department of Computer Science
UNCG
p_manda@uncg.edu

Abhyuday Mandal
University of Georgia
amandal@stat.uga.edu

Breda Munoz
Social, Statistical & Environmental Sciences
RTI International
breda@rti.org

Jerry Reiter
Department of Statistical Science
Duke University
jerry@stat.duke.edu

Scott Richter
Department of Mathematics and Statistics
UNCG
sjricht2@uncg.edu

Ami Shi
SAS
amy.shi@sas.com

Haimeng Zhang
Department of Mathematics and Statistics
UNCG
h_zhang5@uncg.edu

Welcome from the Hosts

On behalf of The University of North Carolina at Greensboro and the North Carolina Chapter of the American Statistical Association, we are excited to welcome you to the 2018 edition of the ***International Conference on Advances in Interdisciplinary Statistics and Combinatorics***, a collaborative project of the NC ASA and UNCG. We thank you for choosing to attend AISC-2018 and sincerely hope that your participation in the conference will be both productive and enjoyable. We hope that you enjoy your time on the UNC Greensboro campus and in the city of Greensboro. If you are joining us from out of state, we hope that during your visit to North Carolina you are able to experience some of NC's unique features – such as Carolina barbeque, local craft breweries, and the beautiful fall colors characteristic of the Appalachian State.

We would like to thank all the external sponsors of this conference including IMA, NISS, Pearson, RHO, SAS, and Springer. This generous support helped us provide partial support to more than 25 young researchers, both from North Carolina as well as beyond. Additionally, we would like to recognize local support from UNC Greensboro including Dr. John Kiss, Dean College of Arts and Sciences, Dr. R. Shivaji, Head, Mathematics and Statistics, and the conference secretaries Haley Childers and Carri Richter who worked tirelessly behind the scene. We also want to thank the very dedicated group of volunteers who have helped tremendously. Similarly support from the North Carolina Chapter of ASA was very critical.

We also want to thank all of the plenary speakers and the session organizers whose contributions enriched the conference program. We would specifically like to thank and acknowledge the five distinguished NC statisticians (David Banks, David Dickey, Sujit Ghosh, Breda Munoz, and Maura Stokes) who have graciously agreed to be honored by the AISC community at the conference banquet Saturday evening.

We trust that you will take away good memories with friends and colleagues at AISC. We welcome you to join the NC ASA Chapter as a member and hope to see you at a future NC ASA event soon.

Sat Gupta – UNC Greensboro
AISC 2018 Conference Chair

Elizabeth Mannshardt
NC State University &
NC-ASA President 2018

Conference Program
AISC 2018: October 5–7, 2018; UNC Greensboro

Please Note: At UNCG Campus, shuttles will always drop off and pick up from Sterling Street between Elliott University Center (EUC: conference venue) and Walker Parking Deck.

October 4, 2018, Thursday

5:30 – 7:30 pm **Registration Desk, Azalea Room at the conference hotel,**
Holiday Inn, Gate City Blvd

October 5, 2018, Friday

8:00 – 8:45 **Shuttles from Holiday Inn to UNCG**
Departures from Holiday Inn: 8:00 am, 8:00 am, 8:15 am, 8:30 am, 8:45 am
Please do not wait for the last shuttle. You may remain stranded and may have to take an Uber.

7:30 – **Registration Desk, Refreshments, EUC Auditorium Lobby**

9:00 – 9:30 **Inaugural Remarks – EUC Auditorium**

Sat Gupta – Conference Chair
R. Shivaji – Head, Mathematics and Statistics, UNCG
Kelly Burke, Vice Provost & Dean Graduate School
Elizabeth Mannshardt, NC State University & NC-ASA President
John Kiss – Dean College of Arts and Sciences, UNCG

9:30 – 10:10 **October 5 – Plenary Session 1 – EUC Auditorium**
Chair: David Banks – Duke University & SAMSI

Barry Nussbaum – ASA President (2017)
It's Not What We Said, It's Not What They Heard, It's What They Say They Heard

10:10 – 10:30 **Coffee Break**

10:30 – 12:30 **October 5 – Parallel Session 1A – Sharpe Room EUC**
Statistical Distributions with Applications in Economics and Health Sciences
Organizer & Chair: Indranil Ghosh – UNC Wilmington

George Yanev – The University of Texas Rio Grande Valley
Borel-Tanner Distribution and Bayes Estimators for the Basic Reproduction Number of an Epidemic

Swati Deb Roy – The University of South Carolina Beaufort
Quantitative Modeling Applied to Childhood Obesity

Rachel M. Carroll – UNC Wilmington
A Bayesian Approach to Weighted Quantile Sums with Extension for Time to Event Data

Dhanamalee K. Bandara – UNC Wilmington
A Neighborhood Hypothesis Test for Functional Data

10:30 – 12:30 **October 5 – Parallel Session 1B – Dail Room EUC**
Sampling Methods in High-Dimensional Spaces
Organizer & Chair: Cheng Cheng – Duke University & SAMSI

Matthias Sachs – Duke University & SAMSI
Efficient Numerical Algorithms for the Generalized Langevin Equation

Cheng Cheng – Duke University & SAMSI
Stable Phaseless Sampling and Reconstruction of Real-valued FRI Signals

Pulong Ma – SAMSI & Duke University
A Fused Gaussian Process Model for Very Large Spatial Data

Qiuyi Wu – Rochester Institute of Technology
Machine Learning for Music Mining with LDA Model

10:30 – 12:30 **October 5 – Parallel Session 1C – Claxton Room EUC**
Nonparametric Methods
Organizer & Chair: Scott Richter – UNC Greensboro

Scott J Richter – UNC Greensboro
Simultaneous Confidence Intervals for Comparing Scale Parameters using Deviances

Melinda McCann – Oklahoma State University
An Efficient Multiplicity Adjustment for Large Scale Chi-Square Endpoints

Sayed Mostafa – North Carolina A&T State University
Finite Population Model-Assisted Estimation using Combined Parametric and Nonparametric Regression Smoothers

Mubbasher Munir – University of Management and Technology
Lahore, Pakistan
The Negative Impact of Outliers in Linear Regression Model; Possible Solutions Using Robust Regression

10:30 – 12:30 **October 5 – Parallel Session 1D – Alexander Room EUC**
Bayesian Nonparametric Methods and Applications
Organizer & Chair: Andres F. Barrientos – Duke University

Shaobo Han – Duke University
Tensor Decomposition for Multiple Spatial Passing Networks

Olanrewaju Akande – Duke University
Simultaneous Edit and Imputation For Household Data With Structural Zeros

Andee Kaplan – Duke University
Counting Casualties in The Syrian Conflict with Bayesian Record Linkage

Andres F. Barrientos – Duke University
A Bayesian Goodness-Of-Fit Test for Regression

10:30 – 12:30 **October 5 – Parallel Session 1E – Kirkland Room EUC**
Sampling Methods
Organizer: Geeta Kalucha – PGDAV College, University of Delhi
Chair: Hina Khan – Government College University, Lahore, Pakistan

G N Singh – IIT (ISM) Dhanbad, India
Improved Estimation Procedures of Population Parameter for Sensitive Characteristic using Randomized Response Technique

Sat Gupta – UNC Greensboro

Randomized Response Techniques for Estimation of Small Area Total

Geeta Kalucha – PGDAV College, University of Delhi
Ratio Estimation of the Mean under RRT Models

Zaheen Khan – Federal Urdu University of Arts, Science and Technology Islamabad, Pakistan
An Optimal Systematic Sampling Scheme

12:30 – 2:00 **Lunch at Moran Commons Dining Center**

2:10 – 3:30 **October 5 – Plenary Session 2 – EUC Auditorium**
Chair: John Stufken – Arizona State University

David Banks – Duke University & SAMSI
Statistical Issues with Agent-Based Models

Mahlet Tadesse – Georgetown University
Variable Selection in Mixture Models: Uncovering Cluster Structure and Relevant Features

3:30 – 3:50 **Coffee Break**

3:50 – 5:50 **October 5 – Parallel Session 2A – Sharpe Room EUC**
Lifetime Data Analysis
Organizer & Chair: Suvra Pal – University of Texas at Arlington

Indranil Ghosh – UNC Wilmington
New Class of Skewed Distributions with Applications in Environmental Science

Rajarshi Dey – University of South Alabama
On Hazard Function of Kumaraswamy Distribution

Sy Han (Steven) Chiou – University of Texas at Dallas
Generalized Scale-Change Models for Recurrent Event Processes Under Informative Censoring

Suvra Pal – University of Texas at Arlington
Destructive Cure Rate Model Based on Multiple Treatments

3:50 – 5:50 **October 5 – Parallel Session 2B – Dail Room EUC**
Advances in Biostatistics and Epidemiology
Organizer & Chair: Sujit Ghosh – NC State University

Brian Neelon – Medical University of South Carolina
Bayesian Zero-Inflated Negative Binomial Regression Based on Pólya-Gamma Mixtures

Rajib Paul – UNC Charlotte
Assessing Health Disparities using Nonparametric Multivariate Density Estimation subject to Marginal Unimodality Constraints

Yunran Chen – Duke University
Testing Poisson versus Poisson mixtures with applications to neuroscience

Shrabanti Chowdhury – Icahn School of Medicine at Mount Sinai
Group Regularization for Zero-Inflated Models with Application to Health Care Demand in Germany

3:50 – 5:50 **October 5 – Parallel Session 2C – Claxton Room EUC**
Undergraduate Research with Data-Driven Analysis

Organizer & Chair: Mark Daniel Ward – Purdue University

Erik Wendt – Gettysburg College
Minimum Entropy Clustering of Functional Data

Siri Neerchal – University of Maryland
Assessing the Descriptive Epidemiology of Idiopathic Clubfoot in Iowa

Amelia Schroeder – East Tennessee State University
Comparison of Machine Learning Algorithms for Rapid Evaporative Ionization Mass Spectrometry (REIMS) Studies

James Marshall Reber – Purdue University
Markov Chains, Mixing Times, and Couplings

Michael A Smith – Purdue University
A Data-Driven Approach to Combinatorial Game Theory

6:00 – 8:00 **Reception Cone Ball Rooms – EUC**
Hosted by Springer

8:00 – 8:45 **Shuttles Back to Holiday Intn**
Departures from outside of EUC: 8:00 pm, 8:15 pm, 8:30 pm, 8:45 pm
Please do not wait for the last shuttle. You may remain stranded and may have to take an Uber.

October 6, 2018, Saturday

8:00 – 8:45 **Shuttles from Holiday Inn to UNCG**
Departures from Holiday Inn: 8:00 am, 8:00 am, 8:15 am, 8:30 am, 8:45 am
Please do not wait for the last shuttle. You may remain stranded and may have to take an Uber.

8:00 – 9:00 **Registration/Refreshments – EUC Auditorium Lobby**

9:00 – 10:20 **October 6 – Plenary Session 3 – EUC Auditorium**
Chair: William Li – Shanghai Advanced Institute of Finance

David Dickey – NC State University
Unit Root Testing, a Historical Perspective

Sujit Ghosh – NC State University
Does Knowledge of Shapes Matter in Statistics?

10:20 – 10:40 **Coffee Break**

10:40 – 12:40 **October 6 – Parallel Session 3A – EUC Auditorium**
Design of Experiments: Factorial Experiments
Organizers: John Stufken (Arizona State University) & Abhyuday Mandal (University of Georgia)
Chair: John Morgan – Virginia Tech

Bobby Mee – University of Tennessee
Utilizing the Block Diagonal Covariance Structure of Nonregular Two-Level Designs

Huaiqing Wu – Iowa State University
Fractional Factorial Designs with Clear Two-Factor Interactions

William Li – Shanghai Advanced Institute of Finance
Individual Clear Effects for Fractional Factorial Designs

Arman Sabbaghi – Purdue University
An Algebra for the Conditional Main Effects Parameterization

10:40 – 12:40 **October 6 – Parallel Session 3B – Sharpe Room EUC**
Advanced Statistical Modeling of Genomic Data
Organizer & Chair: Sunyoung Shin – University of Texas at Dallas

Naim Rashid – University of North Carolina at Chapel Hill
Modeling Between-Study Heterogeneity for Improved Reproducibility in Gene Signature Selection and Clinical Prediction

Rhonda Bacher – University of Florida
SCnorm: A quantile-regression based approach for normalization of single-cell RNA-seq data

Sunyoung Shin – University of Texas at Dallas
atSNP Search: a web resource for statistically evaluating influence of human genetic variation on transcription factor binding

Yuchao Jiang – University of North Carolina at Chapel Hill
Full-spectrum Copy Number Variation Detection by High-throughput DNA Sequencing

Keegan Korthauer – Harvard T.H. Chan School of Public Health & Dana-Farber Cancer Institute
De novo detection and accurate inference of differentially methylated regions

10:40 – 12:40 **October 6 – Parallel Session 3C – Dail Room EUC**
High-Dimensional Variable Selection
Organizer & Chair: Sy Han (Steven) Chiou – University of Texas Dallas

Eliana Christou – University of North Carolina at Charlotte
Central Quantile Subspace

Ana Kenney – Penn State University
Computationally Viable and Effective Feature Selection in L_0 Norm

Thierry Chekouo T – University of Calgary
Bayesian Factor Analysis Regression with Incorporation of Grouping Information

Wenjie Wang – University of Connecticut
Integrative Survival Analysis with Uncertain Event Times in Application to a Suicide Risk Study

10:40 – 12:40 **October 6 – Parallel Session 3D – Claxton Room EUC**
Best Presentation Award Session for Undergraduate Students
Chair: Mark Daniel Ward – Purdue University
Jury: Haimeng Zhang (UNC Greensboro), Ivette Gomes (University of Lisbon), Suvra Pal (University of Texas at Arlington)

Yiwen Tang – Wake Forest University
Classifying Hate Speech Using a Two-Stage Model

Amber Young – Purdue University
Newer Variations of the Unrelated Question Binary RRT Model Examining the Impact of Untruthful Responding

Zavia Epps – Jackson State University
Evaluation of Deaggregation Methods Used to Simulate Unknown Premises Locations for Disease Models: Cattle Disease Outbreaks in Argentina

Austin Miller – University of Wyoming
Robust Trend Filtering and Outlier Detection

Shuchi Goyal – University of Georgia
Hierarchical Bayes Unit-Level Small Area Estimation Model for Normal Mixture Populations

Adrianna Kallis – Iowa State University
Visualizing IPUMS-I Demographics

10:40 – 12:40 **October 6 – Parallel Session 3E – Alexander Room EUC**
Best Presentation Award Session for Graduate Students
Chair: David Banks – Duke University & SAMSI
Jury: Scott Richter (UNC Greensboro), Sujit Ghosh (NC State University), David Banks (Duke University & SAMSI)

Bin Luo – UNC Greensboro
Penalized Re-Descending M Estimation with Non-Convexity for High-Dimensional Asymmetric Data

Joshua Lukemire – Emory University
d-QPSO: A Quantum-Behaved Particle Swarm Technique for Finding D-Optimal Designs with Discrete and Continuous Factors and a Binary Response

Abigael Nachtsheim – Arizona State University
Nonparametric Sub-Sampling for Big Data

Austin Lawson – UNC Greensboro
Persistence Curves: A New Vectorization of Persistence Diagrams

Liu Kai – McMaster University
Semi-Parametric Likelihood Inference for Birnbaum-Saunders Frailty Model

Claire Kiers – UNC Chapel Hill
Hurricanes and Rate-Induced Tipping

12:40 – 2:10 **Lunch at Moran Commons Dining Center**
12:40 – 2:10 **JSTP Editorial Board Luncheon – Moran Commons Dining Center, Separate Reserved Area**

2:15 – 4:15 **October 6 – Parallel Session 4A – EUC Auditorium**
Design of Experiments: Computer Experiments and Beyond
Organizers: John Stufken (Arizona State University) & Abhyuday Mandal (University of Georgia)
Chair: Angela Dean – Ohio State University

Thomas Santner – Ohio State University
Variable Selection for Deterministic Computer Simulator Output

Abhyuday Mandal – University of Georgia
EzGP: Easy-to-Interpret Gaussian Process Models for Computer Experiments with Both Quantitative and Qualitative Factors

Jon Stallings – NC State University
Sequential Design and Analysis of Mixture Experiments based on Gaussian Processes

Tirthankar Dasgupta – Rutgers University

Randomization Based Inference from Unbalanced Split Plot Designs

2:15 – 4:15 **October 6 – Parallel Session 4B – Sharpe Room EUC**
Complex High-Dimensional Data Analysis
Organizer & Chair: Xiaoli Gao – UNC Greensboro

Kei Hirose – Kyushu University, Japan
Prenet Penalization in Factor Analysis and its Applications

Yuan Huang – University of Iowa
A Joint Learning of Multiple Precision Matrices with Sign Consistency

Vadim Sokolov – George Mason University
Deep Learning for Spatio-Temporal Modeling

Jonathan P Williams – UNC Chapel Hill
Non-penalized Variable Selection via Generalized Fiducial Inference

Xiaoli Gao – UNC Greensboro
Robust Stochastic Gradient Decent for Online Learning

2:15 – 4:15 **October 6 – Parallel Session 4C – Dail Room EUC**
Topological Data Analysis
Organizer & Chair: Yu-Min Chung – UNC Greensboro

Edgar Lobaton – NC State University
Localized Topological Metric for Edge Detection in Images using CNNs

Vasileios Maroulas – University of Tennessee
Distributions of Persistence Diagrams and Approximations

Yu-Min Chung – UNCG
Topological Fidelity in Image Thresholding

Sayan Mukherjee – Duke University
How Many Directions Determine a Shape and other Sufficiency Results for Two Topological Transforms

2:15 – 4:15 **October 6 – Parallel Session 4D – Claxton Room EUC**
Recent Results in Combinatorics
Organizer & Chair: Clifford Smyth – UNC Greensboro

John Engbers – Marquette University
Extremal Independent Sets and Colorings in k-Chromatic Graphs

C. Matthew Farmer – UNC Greensboro
The Non-Crossing Bond Lattice

David Galvin – University of Notre Dame
Total Non-negativity of Some Combinatorial Matrices

James Rudzinski – UNC Greensboro
Efficient Generation of Unlabeled Graphs

Clifford Smyth – UNC Greensboro
Combinatorial Formulas for Restricted Stirling and Lah Number Matrices and their Inverses

2:15 – 4:15 **October 6 – Parallel Session 4E – Alexander Room EUC**
Genomics and Health
Organizer & Chair: Sunil Mathur – Texas A&M University-Corpus Christi

Sujay Dutta – University of Akron
Statistical Issues in Analyzing Next-Generation Sequencing Data

Bo Li – The Citadel
Simultaneous Inference of Differentially Expressed Isoforms for RNA Sequencing Data

Prashant Waiker – UNC Greensboro
Using Statistics to solve biological problems: An example of termite recombination

Jianping Sun – UNC Greensboro
Multivariate Association Test for Rare Variant Controlling for Cryptic and Family Relatedness

Sunil Mathur – Texas A&M University-Corpus Christi
Detecting Variability in Genome with Applications to Public Health

4:15 – 4:45 **Coffee Break**

4:45 – 5:45 **October 6 – Special Session: Perspectives on Statistical Consulting – A Panel Discussion**
Organizer: Emily Griffith – NC State University

Panel:
David Dickey – NC State University (Moderator)
Sujit Ghosh – NC State University
Siyun Yang – Duke University
Aric LaBarr – Elder Research
Scott Richter – UNC Greensboro

6:00 – 6:30 **NC-ASA Chapter Meeting (non-members also invited) – EUC Auditorium**

6:30 – 9:00 **Conference Banquet & NC-ASA Chapter Awards Ceremony, EUC Cone Ballrooms**

9:00 – 9:45 **Shuttles Back to Holiday Inn**
Departures from outside of EUC: 9:00 pm, 9:15 pm, 9:30 pm, 9:45 pm
Please do not wait for the last shuttle. You may remain stranded and may have to take an Uber.

October 7, 2018, Sunday

8:00 – 8:45 **Shuttles from Holiday Inn to UNCG**
Departures from Holiday Inn: 8:00 am, 8:15 am, 8:30 am, 8:45 am
Please do not wait for the last shuttle. You may remain stranded and may have to take an Uber.

8:00 – 9:00 **Registration/Refreshments – EUC Auditorium Lobby**

9:00 – 10:20 **October 7 – Plenary Session 4 – EUC Auditorium**
Chair: Javier Rojo – Oregon State University

Christopher Nachtsheim – University of Minnesota
Design Oriented Modeling

Maria Ivette Gomes – University of Lisbon
Generalized Means, Linear Combinations and Bias Reduction in the Estimation of Heavy Tails

10:20 – 10:40 **Coffee Break**

10:40 – 11:20 **October 7 – Plenary Session 5 – EUC Auditorium**
Chair: Maria Ivette Gomes – University of Lisbon

Javier Rojo – Oregon State University
Tails of Distributions – Classification and Testing

11:25 – 1:00 **October 7 – Parallel Sessions 5A – EUC Auditorium**
Design of Experiments: Big Data and More
Organizers: John Stufken (Arizona State University) & Abhyuday Mandal (University of Georgia)
Chair: Abhyuday Mandal – University of Georgia

John Stufken – Arizona State University
Information-Based Subdata Selection

Min Yang – University of Illinois at Chicago
Information-Based Optimal Subdata Selection for Mixture Modelling

Rakhi Singh – IITB-Monash Research Academy
Pseudo Generalized Youden Designs

11:25 – 1:00 **October 7 – Parallel Sessions 5B – Sharpe Room EUC**
Sampling Theory and Methods
Chair: G. N. Singh – IIT (ISM) Dhanbad, India

Javid Shabbir – Quaid-i-Azam University, Pakistan
Use of Successive Sampling Strategy for Finite Population Distribution Function Under Nonresponse

Sadia Khalil – Lahore College for Women University
Mean Estimation of Sensitive Variables under Measurement Errors Using Optional RRT Models

Qi Zhang – UNC Greensboro
Comparison of Mean Estimators of Sensitive Variables under Measurement Errors with Respect to Efficiency and Respondent Privacy

11:25 – 1:00 **October 7 – Parallel Sessions 5C – Dail Room EUC**
Mathematics and Statistics for the Earth's Climate System
Organizer & Chair: Christian Sampson – SAMSI

Christian Sampson – SAMSI
Statistical Topography for Sea Ice Modeling

Colin Guider – UNC Chapel Hill
Ensemble Data Assimilation on a Non-Conservative Adaptive Mesh

Yu-Min Chung – UNC Greensboro
Computational Topology on Sea Ice Data

11:25 – 1:00 **October 7 – Parallel Sessions 5D – Claxton Room EUC**
Health Effects of Environment Pollution (Fri or Sun AM)
Organizer & Chair: Hrishikesh Chakraborty – Duke University

Sohini Raha – NC State University
On the Probability Distribution of Durations of Heatwaves

Yawen Guan – NC State University & SAMSI
Using Mobile Monitors for Fine-Scale Spatiotemporal Air Pollution Analysis

Philip White – Duke University
Modeling Daily Seasonality of Mexico City Ozone using Nonseparable Covariance Models on Circles Cross Time

1:00 – 2:30 **Lunch at Moran Commons Dining Center**

2:30 – 4:30 **October 7 – Parallel Session 6A – EUC Auditorium**
Design of Experiments: Recent Advances
Organizers: John Stufken (Arizona State University) & Abhyuday Mandal (University of Georgia)
Chair: John Stufken – Arizona State University

Lin Wang – UCLA
Optimal Maximin L1-Distance Latin hypercube Designs Based on Good Lattice Point Designs

Ming-Hung (Jason) Kao – Arizona State University
Locally Optimal Designs for Mixed Continuous and Binary Responses

Wei Zheng – University of Tennessee
Design Based Incomplete U-statistics

Xinwei Deng – Virginia Tech
An Efficient Algorithm for I-Optimal Designs of Generalized Linear Models

2:30 – 4:30 **October 7 – Parallel Sessions 6B – Sharpe Room EUC**
Big Data Analytics for Biology, Healthcare, and Science
Organizer & Chair: Somya Mohanty – UNC Greensboro

Somya Mohanty – UNC Greensboro
A Data-Driven Analysis of Patient Rehospitalization Risk

Deborah Lekan – UNC Greensboro
Big Data Analysis of Electronic Healthcare Data

Sayed Mostafa – North Carolina A&T State University
Computing Happiness from Textual Data

Darpan Jhawar – UNC Greensboro
Big Data Analysis of Scientific Publications

Oana Dumitrescu – UNC Greensboro
Statistical Framework for Semantic Similarity Searching on Biological Data

Stacey Miertschein (Winona State University) & Amber Young (Purdue University)
A Data-driven Approach to Predicting Diabetes and Cardiovascular Disease with Machine Learning

2:30 – 4:30 **October 7 – Parallel Sessions 6C – Dail Room EUC**
Sampling Methods
Organizer & Chair: Javid Shabbir – Quaid-i-Azam University, Pakistan

Muhammad Ismail – COMSATS, Lahore, Pakistan
Generalized Ratio-Type Estimator for Population Variance Using Auxiliary Information

in Simple Random Sampling

Hina Khan – Government College Lahore University
Efficient Monitoring of Process Mean Using Exponentially Weighted Moving Average Control Charts, Designed with Ratio Estimators under Ranked Set Sampling

Mahnaz Makhdum – Lahore College for Women University
A Modified Ratio Estimator of Population Mean of a Sensitive Variable in the Presence of Non-Response in Simple Random Sampling

Amber Asghar – NCBA&E Lahore & Virtual University of Pakistan
Regression-cum-Exponential Estimator for Finite Population Variance using Multi-Auxiliary Variables: A Simulation Study

2:30 – 4:30 **October 7 – Parallel Sessions 6D – Claxton Room EUC**
Spatial Statistics
Organizer & Chair: Haimeng Zhang – UNC Greensboro

Haimeng Zhang – UNC Greensboro
Intrinsic Random Functions and Universal Kriging on the Circle

Jong-Min Kim – SAMSI & University of Minnesota at Morris
The Copula Functional ARCH Directional Dependence for Intraday Volatility with High-frequency Financial Data

J. Beleza Sousa – CMA-ISEL Portugal
Stationary Yield to Maturity Zero Coupon Bonds Historical Simulation Value at Risk

Wei Chen – UNC Greensboro
Spectral Density Estimation for Power Model

Jesse Clifton – NC State University
Spatio-Temporal Decisions: Model-Based and Model-Free Approaches

Romesh Ruwan Thanuja – UNC Greensboro
Non-consistency of MOM Variogram Estimators on the Sphere

4:30 – 4:50 **Coffee Break**

5:00 – 5:30 **Shuttles Back to Holiday Inn**
Departures from outside of EUC: 5:00 pm, 5:30 pm
Please do not wait for the last shuttle. You may remain stranded and may have to take an Uber.

Distinguished Service Awards by the NC Chapter of ASA

During the AISC conferences, NC-ASA Chapter honors a small number of NC statisticians for their outstanding contributions to the theory and practice of statistics and/or for their outstanding contributions to the Chapter activities. This year's honorees are:

David Banks

David Banks obtained an M.S. in Applied Mathematics from Virginia Tech in 1982, followed by a Ph.D. in Statistics in 1984. He won an NSF Postdoctoral Research Fellowship in the Mathematical Sciences, which he took at Berkeley. In 1986 he was a visiting assistant lecturer at the University of Cambridge, and then joined the Department of Statistics at Carnegie Mellon in 1987. In 1997 he went to the National Institute of Standards and Technology, then served as chief statistician of the U.S. Department of Transportation, and finally joined the U.S. Food and Drug Administration in 2002. In 2003, he returned to academics at Duke University and is currently the director of the Statistical and Applied Mathematical Sciences Institute.

David Banks was the coordinating editor of the *Journal of the American Statistical Association.* He co-founded the journal *Statistics and Public Policy* and served as its editor. He co-founded the American Statistical Association's Section on National Defense and Homeland Security, and has chaired that section, as well as the sections on Risk Analysis and on Statistical Learning and Data Mining. He has published 74 refereed articles, edited eight books, and written four monographs.

His research areas include models for dynamic networks, dynamic text networks, adversarial risk analysis (i.e., Bayesian behavioral game theory), human rights statistics, agent-based models, forensics, and certain topics in high-dimensional data analysis.

David A. Dickey

Professor Dickey was born and raised in Ohio where he attended Miami University obtaining undergraduate and masters degrees in mathematics. He taught math for 2 years at the College of William and Mary in Virginia and 1 year at Randolph Macon College before returning to school in 1972, earning a PhD at Iowa State University in 1976. There he received the Snedecor Award for top PhD student. His dissertation work on stationarity testing in time series resulted in two well received papers, joint with Wayne Fuller, that solved a long standing problem in Econometrics. According to Google Scholar, August 2018 these papers have received 24,369 and 13,979 citations respectively in the scientific literature.

In 1976 Dr. Dickey accepted a position in Statistics at NC State. He advised or coadvised 16 PhD students and served on a few hundred graduate student advisory committees. He taught graduate level methods, time series, linear models, and data mining courses among others. Dave was a founding faculty member of NC State's Institute for Advanced Analytics in which he taught time series and data mining. He is a member of the NCSU Academy of Outstanding Teachers and the Academy of Outstanding Faculty Engaged in Extension, a result of his statistical consulting activity throughout his 43 year career. Dave is a Fellow of the American Statistical Association and William Neal Reynolds Distinguished Professor. He also teaches advanced statistical short courses for SAS Institute.

Dave is married to Barbara and has 2 grown children: Susan McShane (Ryan) a graphic artist and Michael (Kim) an NCSU Chemical Engineering professor. His grandchildren are Aliyah (9), Emmerson (7), Declan (3), and Gilian (1).

Sujit Kumar Ghosh

Professor Sujit Kumar Ghosh earned a Ph.D. in Statistics from the University of Connecticut in 1996 and is currently a tenured Full Professor in the Department of Statistics at North Carolina State University (NCSU) in Raleigh, NC, USA. He has over 25 years of experience in conducting, applying, evaluating and documenting statistical analysis of biomedical and environmental data. He received the International Indian Statistical Association (IISA) Young Investigator Award in 2008; was elected a Fellow of the American Statistical Association (ASA) in 2009 and was elected as the President of the NC Chapter of ASA in 2013 and also the President of the IISA in 2017. He served as the Program Director in the Division of Mathematical Sciences (DMS) within the Directorate of Mathematical and Physical Sciences (MPS) at NSF in 2013-2014.

Prof. Ghosh has supervised over 35 doctoral graduate students and has published over 100 refereed journal articles in the various areas of statistics with applications in biomedical and environmental sciences, econometrics and engineering. He has also served as a statistical investigator and consultant for over 45 different research projects funded by various leading private industries and federal agencies. In recent years, Prof. Ghosh has contributed significantly in developing statistical models and associated methodologies for various inferential problems that are subject to shape constraint. Most recently, during 2014- 2017 he also served as the Deputy Director at the Statistical and Applied Mathematical Sciences Institute.

A past NC ASA President, Sujit remains a very active member, attending many events. Sujit continues to work with the NC ASA Executive Board via his work as the Core Strategic Team Lead for NC ASA's newly developed Industry Interests Group. Sujit has worked on the IIG's agenda, spearheading implementation of many ideas and initiatives. He is actively working to implement an NC ASA survey of members to gauge interests and needs of the community, and has plans to organize events to meet these needs.

Maura Stokes

Maura Stokes is Senior R & D Director of Statistical Development in the Advanced Analytics Division at SAS Institute and is responsible for the development of the statistical products at SAS, including SAS/STAT software. Her department researches and implements methodology, writes the documentation, and promotes the software with presentations and courses for customer and professional audiences. Previously, she directed testing for the statistical software at SAS, oversaw statistical application development, and served as SAS/STAT development product manager. Lead author of *Categorical Data Analysis with SAS*, she has taught workshops on applied statistical topics for many years. She began her career as a survey statistician for Research Triangle Institute.

Stokes received her DrPH from the Department of Biostatistics at the University of North Carolina. She has held positions in both the American Statistical Association, where she provided leadership in the development of the Conference on Statistical Practice, and the ENAR organization. She has been an adjunct associate professor at the UNC Department of Biostatistics.

Stokes was elected a Fellow of the American Statistical Association in 2008 and received the ASA Founder's Award for distinguished service in 2016.

Breda Munoz

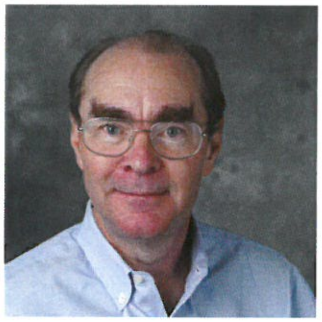
Breda Munoz is a Research Statistician at RTI International. She has more than a dozen years of experience in data management, exploratory data analysis, survey data analysis, and Bayesian data analysis. She also has a wealth of expertise in sampling design, missing data, geostatistics, environmental statistics, clinical trial data analysis, and nonparametric statistics, in addition to statistical consultancy and statistical software programming.

Dr. Munoz is currently focusing her research on nonresponse and missing data analysis, NCD predictive modeling, and the analysis of complex surveys. She is also pursuing the development of models that incorporate information from different sources, such as data from different surveys and studies, geographic information system layers, satellite sensing data, and census data.

Previously, she worked as a research associate at Oregon State University and as a professor at the University of Los Andes and University of Costa Rica. Dr. Munoz is a member of the American Statistical Association, Association of State Wetland Managers, and International Environmetrics Society.

She is a past NC ASA President, and remains an active member. Breda continues to work with the NC ASA Executive Board, organizing events and opportunities for high school statistics students and teachers, providing a vital outreach to an under-served part of North Carolina's statistics community. Breda also served as a Mentor in NC ASA's first mentoring event this spring - an important role in our service to our young professionals.

Plenary Speakers: AISC 2018



David Banks

David Banks obtained an M.S. in Applied Mathematics from Virginia Tech in 1982, followed by a Ph.D. in Statistics in 1984. He won an NSF Postdoctoral Research Fellowship in the Mathematical Sciences, which he took at Berkeley. In 1986 he was a visiting assistant lecturer at the University of Cambridge, and then joined the Department of Statistics at Carnegie Mellon in 1987. In 1997 he went to the National Institute of Standards and Technology, then served as chief statistician of the U.S. Department of Transportation, and finally joined the U.S. Food and Drug Administration in 2002. In 2003, he returned to academics at Duke University and is currently the director of the Statistical and Applied Mathematical Sciences Institute.

David Banks was the coordinating editor of the *Journal of the American Statistical Association.* He co-founded the journal *Statistics and Public Policy* and served as its editor. He co-founded the American Statistical Association's Section on National Defense and Homeland Security, and has chaired that section, as well as the sections on Risk Analysis and on Statistical Learning and Data Mining. He has published 74 refereed articles, edited eight books, and written four monographs.

His research areas include models for dynamic networks, dynamic text networks, adversarial risk analysis (i.e., Bayesian behavioral game theory), human rights statistics, agent-based models, forensics, and certain topics in high-dimensional data analysis.

David A. Dickey

Professor Dickey was born and raised in Ohio where he attended Miami University obtaining undergraduate and masters degrees in mathematics. He taught math for 2 years at the College of William and Mary in Virginia and 1 year at Randolph Macon College before returning to school in 1972, earning a PhD at Iowa State University in 1976. There he received the Snedecor Award for top PhD student. His dissertation work on stationarity testing in time series resulted in two well received papers, joint with Wayne Fuller, that solved a long standing problem in Econometrics. According to Google Scholar, August 2018 these papers have received 24,369 and 13,979 citations respectively in the scientific literature.

In 1976 Dr. Dickey accepted a position in Statistics at NC State. He advised or coadvised 16 PhD students and served on a few hundred graduate student advisory committees. He taught graduate level methods, time series, linear models, and data mining courses among others. Dave was a founding faculty member of NC State's Institute for Advanced Analytics in which he taught time series and data mining. He is a member of the NCSU Academy of Outstanding Teachers and the Academy of Outstanding Faculty Engaged in Extension, a result of his statistical consulting activity throughout his 43 year career. Dave is a Fellow of the American Statistical Association and William Neal Reynolds Distinguished Professor. He also teaches advanced statistical short courses for SAS Institute.

Dave is married to Barbara and has 2 grown children: Susan McShane (Ryan) a graphic artist and Michael (Kim) an NCSU Chemical Engineering professor. His grandchildren are Aliyah (9), Emmerson (7), Declan (3), and Gilian (1).

Sujit Kumar Ghosh

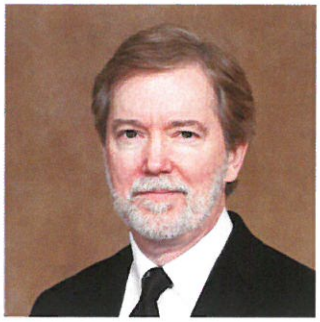
Professor Sujit Kumar Ghosh earned a Ph.D. in Statistics from the University of Connecticut in 1996 and is currently a tenured Full Professor in the Department of Statistics at North Carolina State University (NCSU) in Raleigh, NC, USA. He has over 25 years of experience in conducting, applying, evaluating and documenting statistical analysis of biomedical and environmental data. He received the International Indian Statistical Association (IISA) Young Investigator Award in 2008; was elected a Fellow of the American Statistical Association (ASA) in 2009 and was elected as the President of the NC Chapter of ASA in 2013 and also the President of the IISA in 2017. He served as the Program Director in the Division of Mathematical Sciences (DMS) within the Directorate of Mathematical and Physical Sciences (MPS) at NSF in 2013-2014.

Prof. Ghosh has supervised over 35 doctoral graduate students and has published over 100 refereed journal articles in the various areas of statistics with applications in biomedical and environmental sciences, econometrics and engineering. He has also served as a statistical investigator and consultant for over 45 different research projects funded by various leading private industries and federal agencies. In recent years, Prof. Ghosh has contributed significantly in developing statistical models and associated methodologies for various inferential problems that are subject to shape constraint. Most recently, during 2014- 2017 he also served as the Deputy Director at the Statistical and Applied Mathematical Sciences Institute.

A past NC ASA President, Sujit remains a very active member, attending many events. Sujit continues to work with the NC ASA Executive Board via his work as the Core Strategic Team Lead for NC ASA's newly developed Industry Interests Group. Sujit has worked on the IIG's agenda, spearheading implementation of many ideas and initiatives. He is actively working to implement an NC ASA survey of members to gauge interests and needs of the community, and has plans to organize events to meet these needs.

Maria Ivette Gomes

Ivette Gomes was a Full Professor at the Department of Statistics and Operations Research, Faculty of Sciences, University of Lisbon (1988-2011), being now an Emeritus Professor at University of Lisbon, and principal researcher at the Centre for Statistics and Applications/University of Lisbon (CEA/UL). She has a PhD in Statistics (University of Sheffield, UK, 1978) and a Habilitation Degree in Applied Mathematics (UL, 1982). One of her main areas of research is Statistics of Extremes. She was a founding member of the Portuguese Statistical Society (SPE), member of several scientific Associations, President of SPE (1989-1993) and Vice-President of the International Statistical Institute (ISI) in the period 2015-2017. She has been involved in the organization of several international conferences, including the 56th Session of ISI, 2007. Among other editorial duties, she has been Chief Editor of Revstat, since 2003, and Associate Editor of Extremes since 2007. In May, 2015, she has been elected as a Corresponding Member of the Academy of Sciences of Lisbon.

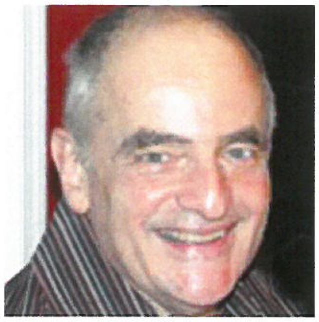


Christopher J. Nachtsheim

Christopher J. Nachtsheim is the Frank A. Donaldson Chair of Operations Management in the Supply Chain and Operations Department of the Carlson School of Management at the University of Minnesota. Dr. Nachtsheim received and his Ph.D. in Operations Research from the University of Minnesota, served as staff member in the Statistics Group at Los Alamos National Laboratory from 1978-1981, and as Senior Statistician at General Mills from 1982-1984. In 1984 he joined the University, serving as Chair in the Department of Operations and Management Science (now Supply Chain and Operations) from 1993-1996 and from 2002-2014, and as Associate Dean of Faculty and Research from 1996-2000.

Dr. Nachtsheim's teaching and research interests center on statistics and business analytics, optimal design of industrial product and process improvement experiments, regression and predictive analytics, and quality management. In addition to his regular teaching duties at the University of Minnesota, Dr. Nachtsheim conducts workshops regularly for industry and has taught business statistics at the Warsaw School of Economics, the Vienna School of Business and Economics, and Lingnan College at Sun Yet Sen University regularly since 1993. Among his major publications are two texts: Applied Linear Statistical Models, 5th Edition, 2005, Richard D. Irwin, and Applied Linear Regression Models, 4th Edition, 2004, Richard D. Irwin (both with John Neter, Michael Kutner, and William Li). He is co-discoverer, along with Brad Jones (SAS Institute) of Definitive Screening Designs, and is the inventor of the coordinate exchange algorithm for constructing exact optimal experimental designs.

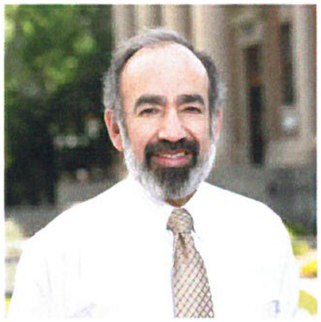
Dr. Nachtsheim is the recipient of the Teacher of the Year Award for the Vienna Executive MBA program in 2017, and the Curtis Cup Award for CEMBA Faculty of the Year in the Carlson Executive MBA program in 2018. Professor Nachtsheim has published over 70 articles in the statistics literature and has served as associate editor for many of the top journals in his field, including *Journal of the American Statistical Association*, *Technometrics*, *Journal of Quality Technology*, *Statistics and Computing*, and *Journal of Statistical Computation & Simulation*. He is currently Associate Editor for the *Journal of Quality Technology*. He served as Examiner, Malcolm Baldrige National Quality Award in 1996. He is a four-time recipient (1991, 2009, 2011, and 2014) of the Brumbaugh Award of the ASQ for best paper published in the area of industrial quality control, two-time recipient of the Lloyd S. Nelson Award of the ASQ for the published paper having the greatest impact on practitioners (2010 and 2012), a two-time recipient of the Jack Youden Prize for the best expository paper published in *Technometrics* (2011 and 2013), and the recipient of the 1992 ASME (CAE/CAD/CAM) National Best Paper Award. Dr. Nachtsheim is a Fellow of the American Statistical Association.



Barry D. Nussbaum

Barry D. Nussbaum was the Chief Statistician for the U.S. Environmental Protection Agency from 2007 until his retirement in March 2016. He started his EPA career in 1975 in mobile sources and was the branch chief for the team that phased lead out of gasoline. Dr. Nussbaum is the founder of the EPA Statistics Users Group. In recognition of his notable accomplishments he was awarded the Environmental Protection Agency's Distinguished Career Service Award.

Dr. Nussbaum has a bachelor's degree from Rensselaer Polytechnic Institute, and both a master's and a doctorate from the George Washington University. In May 2015, he was elected the 112th president of the American Statistical Association. He has been a fellow of the ASA since 2007 and is an elected member of the International Statistical Institute. He has taught graduate statistics courses for George Washington University and Virginia Tech and has even survived two terms as the treasurer of the Ravensworth Elementary School PTA.



Javier Rojo

Javier received his Ph.D. in Statistics from the University of California at Berkeley under the direction of Erich L. Lehmann. He is currently the Korvis Professor of Statistics at Oregon State University. Prior to that, he was the Seneca C. and Mary B. Weeks endowed Chair of Statistics and Chair of the Department of Mathematics and Statistics at the University of Nevada at Reno. Prior to that he was Professor of Statistics at Rice University 2001-2013, and before that he was assistant, associate, and then full professor in the Department of Mathematical Sciences at the University of Texas, El Paso 1984-2001, where he was also the founding director of the BioStatistical Laboratory.

Javier is an elected Fellow of the following societies: American Statistical Association, The Institute of Mathematical Statistics, The Royal Statistical Society, The American Association for the Advancement of Science, and is an elected member of the International Statistical Institute. He also received the 2010 Don Owen award from the American Statistical Association, and recently received the 2018 Etta Z. Falconer award "*In recognition of your outstanding contributions to diversifying the landscape of the mathematical and statistical sciences through excellence, mentorship and leadership.*" He served (1998-1999) as program director in the statistics and probability programs in NSF and has participated as a member and as chair of two subcommittees in the 2013 NSF-DMS Committee of Visitors. He again served in the NSF-DMS Committee of Visitors in 2016.

Additionally, Javier is the Director of the successful REU program (RUSIS@OSU). More than two hundred and thirty three students have participated in the program. A substantial percentage of these students have pursued graduate degrees in the mathematical and statistical sciences. Over thirty-three RUSIS alumni have received Ph.D. degrees in the Mathematical and Statistical Sciences from some of the top universities and a similar number of RUSIS alumni are currently pursuing the doctoral degree. The American Mathematical Society recognized RUSIS in 2014 with the award "Programs that make a difference" in recognition of the impact of the program in mentoring of students and encouraging them to pursue graduate studies. About 61% of the participants are members of underrepresented and nontraditional groups in the mathematical and statistical sciences.

Javier was the Editor of Erich L. Lehmann's Selected works published by Springer in 2012 and of several IMS LNMS volumes on the Lehmann Symposia on Optimality. He is a member of SAMSI's National Advisory Committee and currently serves on the editorial boards of the Journal of Nonparametric Statistics and Involve.

Abstracts of the talks

On the Use of Randomized Response Technique for Improving Mean Estimation of Non-Sensitive Variables in Sensitive Domains

Shakeel Ahmed
Department of Statistics Quaid-i-Azam University Islamabad, Pakistan
sahmed@stat.qau.edu.pk
Coauthors: Javid Shabbir, Department of Statistics Quaid-i-Azam University Islamabad, Pakistan

In social surveys, researchers often have interest in obtaining separate estimates for the parameters of the study variable in specific domains, categorized according to some grouping such as political affiliation, income status, and ethical affiliations. When complete information about domain membership is not available in advance, surveyors need to include questions asking about their domain membership during the survey. The response about membership can easily be collected for non-sensitive domains, but not so for sensitive domains. The respondent may refuse to expose their membership, or provide evasive response about membership in sensitive domains. In such cases a reliable estimate for the mean or total of a non-sensitive quantitative variable in specific domains is quite difficult as we have no truthful information about the domain membership. In such situations, we are either left with an unreliable estimate for small area parameters due to smaller sample size, or get biased estimates following false responses from respondents. To reduce such biases, and to increase response rate on domain membership and to obtain a reliable estimate for the mean of a non-sensitive quantitative study variable in sensitive domains, we use Warner's randomized response technique. Finite sample properties of the proposed estimators are studied. We also compare the proposed estimator with the directly obtained mean estimator for specific domain in terms of efficiency and privacy protection.

Simultaneous Edit and Imputation For Household Data with Structural Zeros

Olanrewaju Akande
Duke University
olanrewaju.akande@duke.edu
Coauthors: Andrés Barrientos and Jerome P. Reiter

Multivariate categorical data nested within households often include reported values that fail edit constraints—for example, a participating household reports a child's age as older than his biological parent's age—as well as missing values. Generally, agencies prefer datasets to be free from erroneous or missing values before analyzing them or disseminating them to secondary data users. We present a model-based engine for editing and imputation of household data based on a Bayesian hierarchical model that includes (i) a nested data Dirichlet process mixture of products of multinomial distributions as the model for the true latent values of the data, truncated to allow only households that satisfy all edit constraints, (ii) a model for the location of errors, and (iii) a reporting model for the observed responses in error. The approach propagates uncertainty due to unknown locations of errors and missing values, generates plausible datasets that satisfy all edit constraints, and can preserve multivariate relationships within and across individuals in the same household. We illustrate the approach using data from the 2012 American Community Survey.

Regression-cum-Exponential Estimator for Finite Population Variance using Multi-auxiliary Variables: A Simulation Study

Amber Asghar
Department of Mathematics and Statistics, NCBA&E Lahore, Pakistan & Virtual University of Pakistan
zukhruf10@gmail.com
Coauthors: Russell J. Steele, Department of Mathematics and Statistics, Faculty of Science, McGill University, Canada; Dr. Muhammad Hanif, Department of Mathematics and Statistics, NCBA&E Lahore Pakistan; Dr. Aamir Sanaullah, Department of Mathematics and Statistics, COMSATS Institute of Information Technology Lahore, Pakistan

In this study, we propose a regression-cum-exponential type estimator for estimating a population variance. In particular, unknown population variance estimation is discussed using multi-auxiliary variables in two-phase sampling and different cases are also derived. We also compare the asymptotic properties of existing approaches those of the proposed estimator. Finally, we use a simulation study for our proposed estimator using multi-auxiliary variables to demonstrate the performance of the estimators in finite samples.

SCnorm: A quantile-regression based approach for normalization of single-cell RNA-seq data

Rhonda Bacher
University of Florida
rbacher@ufl.edu
Coauthors: Li-Fang Chu, Ning Leng, Audrey P Gasch, James A Thomson, Ron M Stewart, Michael Newton, and Christina Kendziorski

Single cell RNA-sequencing (scRNA-seq) is a promising tool that facilitates study of the transcriptome at the resolution of a single cell. However, along with the many advantages of scRNA-seq come technical artifacts not observed in bulk RNA-seq studies. The normalization methods traditionally used in bulk RNA-seq were not designed to accommodate these features and, consequently, applying them to the single-cell setting re- sults in artifacts that bias downstream analyses. To address this, we developed SCnorm to enable efficient and accurate scRNA-seq normalization. Simulation and case study results demonstrate that SCnorm pro- vides for increased accuracy in fold-change estimation as well as improvements in downstream inference.

A Neighborhood Hypothesis Test for Functional Data

Dhanamalee Bandara
University of North Carolina Wilmington
bandarad@uncw.edu
Coauthors: Souparno Ghosh, Leif Ellingson, Raziur Rahman, Ranadip Pal

A common problem arising when analyzing high-dimensional or functional data is that estimates of the covariance are not of full rank, resulting in the inverse being degenerate. Munk et al. (2008) applied the idea of a neighborhood hypothesis test to the one- and multi-sample problems for functional data by deriving a test statistic to determine whether a group of means are approximately equal. More precisely, they tested whether the means were within a predetermined distance to each other. Unfortunately, in many applications, this pre-determined distance is difficult to both specify and interpret. In this presentation, we present a modified test for determining whether the distance between a mean and a hypothesized function is less than a proportion of the total population variance. We will derive a test statistic that is asymptotically normal, and present both simulation studies of the power of the procedure and an application to a data set arising from biology.

Statistical Issues with Agent-Based Models

David Banks
Duke University/SAMSI
banks@stat.duke.edu

Agent-based models have become an ubiquitous tool in many disciplines. But too little is known about their statistical properties. This talk reviews the work that has been done in this area, and describes two strategies for improving model fitting and inference. It also attempts to place agent-based modeling within the span of modern Bayesian inference.

A Bayesian Goodness-of-Fit Test for Regression

Andres F. Barrientos
Duke University
andres.barrientos@duke.edu
Coauthors: Antonio Canale

Regression models are one of the most widely used statistical procedures and the validation of their assumptions plays a crucial role during the data analysis process. Unfortunately, this validation usually relies on the availability of tests tailored to the specific model of interest. In this work, we present a novel Bayesian approach to perform hypothesis testing of goodness-of-fit for a broad class of regression models whose response variable is univariate and continuous. We base our proposal on a suitable transformation of the response variable and a Bayesian prior induced by a predictor-dependent mixture model. We perform hypothesis testing via the Bayes factor and discuss its asymptotic properties. The proposed method is implemented by means of a Markov Chain Monte Carlo algorithm. Simulated and real datasets, illustrate the performance of the proposed approach.

A Bayesian Approach to Weighted Quantile Sum Regression with Extension for Time to Event Data

Rachel Carroll
UNCW
carrollr@uncw.edu
Coauthors: Alexandra White Shanshan Zhao

The term environmental mixture is used to describe chemicals that exist together in an environment, usually with high correlation. Weighted quantile sum regression is one of several methods developed specifically for describing environmental mixtures. This method is valuable because it offers a single weighted sum to describe the entire mixture where the weights indicate the chemical contribution to the mixture for the outcome of interest. However, despite being written to a fairly user friendly R package, this method has not been widely adopted by statisticians and epidemiologists. Some of the perceived negatives include: only assessing the mixture effect in a single direction, estimating the weights via a lesser known numeric optimization algorithm, and only being available for linear and logistic regression. The method proposed here employs a Bayesian approach to avoid the computational algorithm as well as an extension for time-to-event outcomes. This new method is compared to the existing method in a simulation study with low, medium, and high correlated mixtures as well as a real data case study and assessed for computational feasibility as well as recovery of the truth laid out in the simulation setting. The results of these comparisons suggest that the Bayesian extension improves upon the results generated in linear and logistic regression and furnishes good results for the survival analysis setting. Computationally, the Bayesian methodology improves upon the numeric optimization algorithm, particularly for highly correlated mixtures. Therefore, the Bayesian paradigm offers an ideal setting for weighted quantile sum regression and its extensions.

Effects of Progressive Physiotherapy along with or without active rest at hospital for the patients with chronic nonspecific Low Back Pain (Lbp) in Bangladesh.

Sonjit Kumar Chakrovorty
Specialised Physiotherapy And Arthritis Research Center(S P HOSPITAL) Dhaka.
sphddhaka@gmail.com
Coauthors: Md. Feroz Kabir 2 Dr. Md. Shahjahan3

Low Back pain (LBP) is a significant health condition globally suffered by the middle and older aged population due to its impact on work disability, absenteeism and cost. Episodes of LBP may be acute, sub-acute, or chronic depending on the duration. Physiotherapy with active rest is one of the choices of treatment as conservative management of LBP. The objective of this study is to find out the effects of progressive physiotherapy along with or without 3 weeks of active bed rest in hospital for the patients with chronic nonspecific LBP in Bangladesh. We randomly selected 20 chronic nonspecific LBP patients 55 years of age who were willing to participate in the study. Blinded Pre and post evaluation was completed using Numeric Pain Rating Scale (NPRS) for pain level and the Roland-Morris LBP and disability questionnaire for disability. We found that the progressive physiotherapy along with 3 weeks of active bed rest in hospital was significantly better than the progressive physiotherapy along without 3 weeks of active bed rest in hospital on NPRS (P=0.01) and the Roland-Morris LBP measures (P=0.01).

Spectral Density Estimation for Power Model

Wei Chen
UNCG, Department of Mathematics and Statistics
w_chen8@uncg.edu

Estimation of the spectral density function is one of the most important research areas in spatial statistics. When the underlying process is stationary, the estimation of spectral density though Periodogram has been well developed, and when the underlying process is intrinsically stationary, the spectral estimation through Variogram is often used. In this paper, we consider the spectral estimation for Power Model. We investigate the properties of the proposed Method of Moments structure estimator, through which the spectral density function estimation is obtained. Simulation studies are conducted to validate our proposed estimation method.

Testing Poisson versus Poisson mixtures with applications to neuroscience

Yunran Chen
Duke University
yc304@duke.edu
Coauthors: Surya Tapas Tokdar

The simplest assumption on neural spike counts would be Poisson distribution. However, non-Poisson behavior is to be expected and has been documented under many situations. It is desirable to test the Poisson hypothesis to make sure the suitability of statistical analysis based on Poisson assumption. Poisson mixtures offer a rich class of alternatives to the Poisson assumption. Traditional testing procedure, such as chi-square test, performs poorly on distinguishing Poisson against Poisson mixtures. Based on this, we proposed testing using Predictive recursion marginal likelihood (PRML) algorithm, a fast recursive algorithm providing an accurate and computationally efficient estimation of mixing densities in mixture models. Extensive comparisons were conducted, showing great improvement in performance of PRML testing over chi-square testing, as measured by ROC-AUC. Furthermore, PRML testing enable testing between different types of Poisson mixtures which may be exhibited by neural spike count records. We consider an application on the multiplexing problem in neuroscience to explore the fluctuation pattern for a single neuron encoding simultaneous stimuli.

Stable Phaseless Sampling and Reconstruction of Real-valued FRI signals

Cheng Cheng
Duke University/ SAMSI
cheng87@math.duke.edu
Coauthors: Qiyu Sun

In this talk, we consider the stable reconstruction of real-valued signals with finite rate of innovations (FRI), up to a sign, from their magnitude measurements on the whole domain or their phaseless samples on a discrete subset. FRI signals appear in many engineering applications such as magnetic resonance spectrum, ultra wide-band communication and electrocardiogram. For an FRI signal, we introduce an undirected graph to describe its topological structure. We establish the equivalence between the graph connectivity and phase retrievability of FRI signals, and we apply the graph connected component decomposition to find all FRI signals that have the same magnitude measurements as the original FRI signal has. We also propose a stable algorithm with linear complexity to reconstruct FRI signals from their phaseless samples on the above phaseless sampling set.

Generalized Scale-Change Models for Recurrent Event Processes under Informative Censoring

Sy Han Chiou
University of Texas at Dallas
schiou@utdallas.edu
Coauthors: Gongjun Xu, Jun Yan, Kieren Marr, and Chiung-Yu Huang

Two major challenges arise in regression analyses of recurrent event data: first, popular existing models, such as the Cox-type models, may not fully capture the covariate effects on the underlying recurrent event process; second, the censoring time remains informative about the risk of experiencing recurrent events after accounting for covariates. We tackle both challenges by a general class of semiparametric scale-change models that allows a scale-change covariate effect as well as a multiplicative covariate effect. The proposed model is flexible and nests several existing models, including the popular proportional rates model, the accelerated mean model, and the accelerated rate model. Moreover, it accommodates informative censoring through a subject-level latent frailty whose distribution is left unspecified. A robust estimation procedure which requires neither a parametric assumption on the distribution of the frailty nor a Poisson assumption on the recurrent event process is proposed to estimate the model parameters. The asymptotic properties of the resulting estimator are established, with the asymptotic variance estimated from a novel resampling approach. As a byproduct, the structure of the model provides a model selection approach among the submodels via hypothesis testing of model parameters. Numerical studies show that the proposed estimator and the model selection procedure perform well under both noninformative and informative censoring scenarios. The methods are applied to data from two transplant cohorts to study the risk of infections after transplantation.

Group Regularization for Zero-Inflated Models with Application to Health Care Demand in Germany

Shrabanti Chowdhury
Icahn School of Medicine at Mount Sinai
shrabanti.chowdhury@mssm.edu
Coauthors: Saptarshi Chatterjee, Himel Mallick, Prithish Banerjee, Broti Garai

In many biomedical applications, covariates are naturally grouped, with variables in the same group being systematically related or statistically correlated. Under such settings, variable selection must be conducted at both group and individual variable levels. Motivated by the widespread availability of zero-inflated count outcomes and grouped covariates in many practical applications, we consider group regularization for Zero-inflated Negative Binomial (ZINB) regression models. Using a least squares approximation of the mixture likelihood and a variety of group-wise penalties on the coefficients, we propose a unified algorithm

(Google: Group Regularization for Zero-inflated Count Regression Models) to efficiently compute the entire regularization path of the estimators. We investigate the finite sample performance of these methods through extensive simulation experiments and the analysis of a German Healthcare demand data set. Finally, we derive theoretical properties of these methods under reasonable assumptions, which further provides deeper insight into the asymptotic behavior of these approaches.

Central Quantile Subspace

Eliana Christou
University of North Carolina at Charlotte
echris15@uncc.edu

Dimension reduction is a useful technique when working with high-dimensional predictors, where straightforward graphical analysis is not possible. Existing dimension reduction techniques focus on the conditional distribution of the response given the covariates, where specific interest focuses on statistical functionals of the distribution, such as the conditional mean, conditional variance and conditional quantile. We introduce a new method for inferring about the conditional quantile of the response given the covariates and we estimate, for a given τ, the τth Central Quantile Subspace (τ-CQS). The purpose of this paper is threefold. First, we focus on cases where the τth conditional quantile depends on the predictor X only through a single linear combination $B'\tau X$ and we show that we can estimate $B\tau$ consistently up to a multiplicative scalar, even though the estimate might be based on a misspecified link function. Second, we extend the result to tth conditional quantiles that depend on the predictor x through a $d\tau$-dimensional linear combination $B'\tau X$, where $B\tau$ is a $p \times d\tau$ matrix, $d\tau > 1$, and propose an iterative procedure to produce more vectors in the τ-CQS, which are shown to be root n consistent. Third and last, we extend our proposed methodology by considering any statistical functional of the conditional distribution and estimate the fewest linear combinations of X that contain all the information on that functional.

Computational Topology on Sea Ice Data

Yu-Min Chung
University of North Carolina at Greensboro
y_chung2@uncg.edu
Coauthors: Christian Sampson

The fluid permeability of sea ice governs a broad range of physical and biological processes in the polar marine environment. For example, in the Arctic, melt pond drainage is largely controlled by the fluid permeability of the ice. Melt ponds in turn have a significant effect on ice albedo, a critical parameter in climate models. The permeability of sea ice depends on its crystallographic type, granular or columnar ice. We employ methods in the field of computational topology to study x-ray tomographic images of sea ice microstructures and classify their differences. This work may be used in the development of stochastic models of sea ice microstructure suitable for use in larger scale sea ice models.

Topological Fidelity in Image Thresholding

Yu-Min Chung
University of North Carolina at Greensboro
y_chung2@uncg.edu
Coauthors: Sarah Day

An automated image thresholding method based on the persistent homology is presented. The primary difference among traditional methods is that the resultant binary image respects underlying topological features. Furthermore, in the presence of noise, the method provides more information to obtain a better estimate of the Betti numbers. Finally, we will show applications to practical datasets – binary alloy from Material Science, and firn from Climatology.

Spatio-temporal decisions: Model-based and model-free approaches

Jesse Clifton
North Carolina State University
jclifto@ncsu.edu

Spatio-temporal decision problems, such as controlling the spread of disease, present an important statistical and computational challenge. We present ongoing work on the estimation of optimal treatment strategies in this setting. In particular, we compare the performance of model-based and model-free reinforcement-learning algorithms under different degrees of model misspecification in simulation experiments based on two real-world disease control problems: white-nosed bat syndrome, and the 2013-2015 Ebola epidemic. We close by discussing directions for the optimal combination of model-based and model-free approaches, as well as optimal exploration in short-time-horizon problems such as these.

Nonparametric predictive inference for reproducibility of tests - recent results

Frank Coolen
Durham University (UK)
frank.coolen@durham.ac.uk
Coauthors: Tahani Coolen-Maturi, Fatimah Alghamdi, Andrea Simkus (all Durham), Filipe Marques (Lisbon)

In recent years reproducibility of hypothesis tests has received increasing interest in statistical practice: If we were to repeat a test under the same circumstances, would we get the same conclusion with regard to rejectance of the null hypothesis? We have developed a nonparametric predictive inference approach to answer this question. This is a frequentist method based on few model assumptions, enabled by the use of imprecise probabilities. At AISC 2016 an introductory overview was presented. We now present recent results, including reproducibility of tests for data collected by the use of randomized response techniques and for likelihood ratio tests. We also discuss some practical considerations of our methods, resulting from initial explorations of their use for tests in pharmaceutical product development,

Randomization Based Inference from Unbalanced Split Plot Designs

Tirthankar Dasgupta
Rutgers University
tirthankar.dasgupta@rutgers.edu
Coauthors: Rahul Mukerjee

Practical considerations often warrant the use of split-plot designs which involve whole-plots that are not of the same size. This paper investigates randomization based causal inference in unbalanced split-plot designs. Two approaches for estimation of the sampling variance of the natural unbiased estimator of a typical treatment contrast are developed and discussed. The issue of minimaxity, with a view to controlling the bias in variance estimation, is also discussed.

Statistical Issues in Analyzing Next-Generation Sequencing Data

Sujay Datta
Dept. of Statistics, University of Akron
sd85@uakron.edu

Over the past two decades, scientific investigation of the systems biology of a living cell has largely been enabled by technological advancements in genomics, proteomics and metabolomics. High-throughput genomic technologies, starting with the microarrays, have made it possible to simultaneously analyze tens of thousands of genes in an organism's genome. In the last few years, the advent of next-generation sequencing (or deep sequencing) has opened a whole new avenue in high-throughput genomics by increasing the coverage, the resolution and the statistical power of such analyses. The RNA-seq technology offers unprecedented information about the transcriptome, but harnessing this information and extracting knowledge from it using bioinformatics tools remain fraught with challenges and present a bottleneck. A great deal of statistical research is now being devoted to this new, interdisciplinary area, resulting in novel methods to extract signals from noisy data and compare signals across multiple experimental conditions. Here we provide a brief yet informative overview of the type of data produced by the NGS technology, the quantitative issues involved, the associated statistical challenges and the methodologies that address those challenges.

Association Between BMI, Race, Age and Gender to Food Choice in South Carolina's Corridor of Shame

Swati DebRoy
University of South Carolina Beaufort
sdebroy@uscb.edu
Coauthors: Valerie Muehleman, Lydia Breland, Alan Warren

Hardeeville-Ridgeland Middle School (HRMS) of 555 students is disadvantaged racially (90% minority) and socioeconomically (85% free or reduced lunch eligible). This study aims to inform the school wellness policy if providing a salad-bar as another food option at lunch will benefit the 51% children who are overweight or obese. For the very first time a salad-bar was introduced to the existing lunch program at HRMS. Data were gathered on students' BMI, race, gender and socioeconomic status at the beginning and end of the 2016-17 school year. Each student ID was matched to their daily food choice. At the midpoint of the academic year a healthy-lifestyle educational campaign was conducted. This talk will present the results of statistical analysis of this research and also implication of the findings in statistical and mathematical modeling of childhood obesity.

An Efficient Algorithm for I-Optimal Designs of Generalized Linear Models

Xinwei Deng
Department of Statistics, Virginia Tech
xdeng@vt.edu
Coauthors: Yiou Li, Department of Mathematical Sciences, DePaul University

The design issues of generalized linear model are undoubtedly challenging. The state-of-the-art works mostly apply to design criteria on the estimates of regression coefficients. It is of great importance to study optimal designs for generalized linear models from the prediction aspects. In this work, we propose a prediction-oriented design criterion, I-optimality, and develop an efficient sequential algorithm of constructing I-optimal designs for generalized linear models. Through establishing the General Equivalence Theorem of the I-optimality for generalized linear models, we obtain an insightful understanding for the proposed algorithm on how to sequentially choose the support points and update the weights of support points of the design. The proposed algorithm is computationally efficient with guaranteed convergence property. Numerical examples are conducted to evaluate the feasibility and computational efficiency of the proposed algorithm.

On Hazard function of Kumaraswamy Distribution

Rajarshi Dey
University of South Alabama
rajarshidey@southalabama.edu
Coauthors: Nutan Mishra

In this presentation, we discuss shape of hazard function of Kumaraswamy distribution. Specifically we establish that the hazard function of Kumaraswamy distribution is either of a bath-tub shape or is increasing. In reliability and survival analysis, it is often of interest to determine the point at which hazard function reaches its minimum when the shape of the hazard function is bath-tub. We propose different estimators of that change point using different methods including maximum likelihood and quantile method and then evaluate their performances.

Unit Root Testing, a Historical Perspective

David A. Dickey
NC State University
dickey@stat.ncsu.edu

This talk is applied in nature, motivating the need to distinguish nonstationary or unit root processes from stationary mean reverting processes. It looks back to computing as it was in the mid 1970s and its use in the development of a hypothesis test to address this issue. Deciding about stationarity is a critical part of most time series studies and is a key to studying many fundamental economic hypotheses. While this is an older topic and available in many texts now, it seems to remain a tool of interest and is in common use. Google Scholar lists over 36000 citations of the initial 2 papers on the subject by Dickey and Fuller. The paper will focus on motivation and illustrative examples.

A statistical framework for ranking semantic similarity searches powered by ontologies.

Oana Dumitrescu
Department of Computer Science, UNCG
o_dumitr@uncg.edu
Coauthors: Alexander Hahn, Xiaoli Gao, Prashanti Manda

Ontologies, and semantic databases annotated using ontologies, are increasing in number, size, and complexity, particularly in the life sciences. This growth necessitates the development of scalable approaches for reasoning over semantic databases in which large stores of data are annotated using multiple, large ontologies and also the development of statistical approaches for evaluating the large numbers of approximate matches that will results from a semantic similarity search. Here, we address these challenges in the context of the Phenoscape Knowledgebase, a semantic database containing over 320589 phenotypes from genetic studies and the evolutionary biology literature, annotated using terms from multiple, large ontologies.

We use semantic similarity scores computed over simulated datasets created from the Phenoscape Knowledgebase to build multi-variate regression models that can predict the baseline of similarity to be expected due to random chance. We use the extreme value distribution in conjunction with the regression to predict an Expect score for each semantic similarity score. The Expect value represents the number of matches that can be expected at the given similarity or higher for a database of a particular size. In addition to absolute similarity scores, these Expect values enable biologists to understand the significance of a given semantic similarity match.

Extremal Independent Sets and Colorings in k-Chromatic Graphs

John Engbers
Marquette University
john.engbers@marquette.edu

Given a family of graphs, which graph in the family has the most number of proper colorings (vertex colorings where adjacent vertices receive different colors)? Tomescu answered this question for n-vertex k-chromatic graphs, and conjectured an answer for n-vertex k-chromatic connected graphs. Recently, Fox, He, and Manners have proved the conjecture for k colors, and Knox and Mojar announced a forthcoming proof of the conjecture for an arbitrary number of colors.

A color class in a proper coloring forms an independent set of vertices, or set of pairwise non-adjacent vertices. Which graph in a family of graphs has the most number of independent sets? We present some results in the family of n-vertex k-chromatic graphs with several different connectivity requirements. The work is in part joint with Aysel Erey, Lauren Keough, and Taylor Short.

Evaluation of Deaggregation Methods Used to Simulate Unknown Premises Locations for Disease Models: Cattle Disease Outbreaks in Argentina

Zavia Epps
Jackson State University
piarricaze@ymail.com

This project focuses on the study of the 2001 foot-mouth-disease (FMD) outbreak of cattle in Argentina. The large scale of concentration is the utilization of a sample population, with the application of clustering algorithms for recreation of a data realization that can be used for aggregating and deaggregating farm sizes and farm locations. Within the Argentina dataset, the exact premises locations are unknown. The demographic information is aggregated at a Partido level, and infected premises (IP) are aggregated using a grid system. The grid system of the infected premises is smaller than the Partido level used to aggregate the demographic data. The aggregated IP grid reports the number of premises in the grid and the locations of the grid cells, but not the exact locations of premises in the grid. The development of the deaggregation methodallows for the demographic and infection data to be deaggregated, through the creation of data sets with estimates of exact farm locations, that can be used in disease modeling.

The Non-Crossing Bond Lattice

C. Matthew Farmer
The University of North Carolina at Greensboro
cmfarmer@uncg.edu
Coauthors: Dr. Joshua Hallam

Let G be a graph with a finite vertex set and edge set. A bond of G is a spanning subgraph of G whose connected components are induced. This collection of bonds form a partially ordered set which is also a lattice. This lattice has what is known as an ER-labeling. We explore a new subposet of this lattice which we call the "non-crossing bond poset" for all graphs finite graphs. Then we aim to show when this subposet ER-labeling and when it does not. This presentation will focus on what the bond lattice of a graph is, examples of when the non-crossing bond poset has a particular ER-labeling and when it does not, and the classification of all graphs whose non-crossing bond poset has an ER-labeling.

Total Non-negativity of Some Combinatorial Matrices

David Galvin
University of Notre Dame
dgalvin1@nd.edu
Coauthors: David Galvin & Adrian Pacurar

Many combinatorial matrices — such as those of binomial coefficients, Stirling numbers of both kinds, and Lah numbers — are known to be totally non-negative, meaning that all minors (determinants of square submatrices) are non-negative.

The examples noted above can be placed in a common framework: for each one there is a non-decreasing sequence $(a_1, a_2, \ldots)$, and a sequence $(e_1, e_2, \ldots)$, such that the (m, k)-entry of the matrix is the coefficient of the polynomial $(x - a_1) \cdots (x - a_k)$ in the expansion of $(x - e_1) \cdots (x - e_m)$ as a linear combination of the polynomials $1, x - a_1, \ldots, (x - a_1) \cdots (x - a_m)$.

We consider this general framework. For a non-decreasing sequence $(a_1, a_2, \ldots)$ we establish necessary and sufficient conditions on the sequence $(e_1, e_2, \ldots)$ for the corresponding matrix to be totally non-negative. As an application we obtain total non-negativity of a family of matrices associated with chordal graphs.

Robust Stochastic Gradient Decent for Online Learning

Xiaoli Gao
University of North Carolina at Greensboro
x_gao2@uncg.edu

Stochastic gradient descent (SGD) has become a popular approach for online learning when data arrives in a stream or data sizes are very large. Since SGD method obtains the final estimates by adding only one data point at a time and recursively updates the parameter estimate, it has numerical convenience and memory efficiency in Big-data analysis. However, the recursive updating can be inefficient and the SGD estimate is damaged when the data contamination occurs at any recursive step. In this paper, we propose a robust procedure for stochastic gradient descent, which, upon the arrival of each observation, updates the SGD estimates as well as quantifying the outlying possibility of this observation. The proposed method is easy to implement in practice. The finite-sample performance and numerical utility is evaluated by both simulation studies and real data applications.

New Class of Skewed Distributions with Applications in Environmental Science

Indranil Ghosh
University of North Carolina, Wilmington (Department of Mathematics and Statistics)
GHOSHI@UNCW.EDU
Coauthors: Prof. H.K.T.Ng (Southern Methodist University, Department of Statistical Science)

In this paper, we generate a new class of skewed distributions by invoking arguments as described by Ferreira et al. (2006). In particular we consider a specific univariate, absolutely continuous sym- metric distribution, namely, the logistic distribution. Next, using the logistic kernel, we derive a new univariate distribution, henceforth will be known as truncated-logistic skew symmetric distribution (or in short TLSS, wherever they appear in this article). We provide some structural properties of the newly developed distribution. A small simulation study is conducted to study the efficacy of the maximum likelihood method to estimate to model parameters. For illustrative purposes, a real life data set is used to exhibit the applicability of such a model.

Does Knowledge of Shapes Matter in Statistics?

Sujit K Ghosh
NC State University
sujit.ghosh@ncsu.edu

In many practical scenarios, from astronomy to zoology, often there's scientific knowledge that inform us about underlying shape of a curve relating two objects or the distribution of objects in a given population. A few examples include mass-radius relations between exoplanets in astronomy are known to preseve monotone relations, projectiles of objects are known to follow a concave path, production and utility theory in economics prescribes various convexity constraint, dose-response curves are monotone with asymptote, densities of log-returns in finance are unimodal but non-normal. This talk provides a brief tour of shape constrained methodologies for regression and density estimation and explores whether and how the knowldge of knowing shapes of objects may inform statistical inference. Several computational algorithms are presented and also a few unsolved problems are posed as challenges.

Generalized Means, Linear Combinations and Bias Reduction in the Estimation of Heavy Tails

M. Ivette Gomes
CEAUL and DEIO, Faculty of Science, University of Lisbon
ivette.gomes@fc.ul.pt

Heavy-tailed data, from an underlying model with a Pareto-type right tail function, are quite common in financial and telecommunication traffic time series, among other areas of application. When analyzing such a type of data one never knows how much the underlying model differs from a strict Pareto model. And this is the unique situation where the Hill estimator of a positive *extreme value index* (EVI) is "perfect". Asymptotically best linear bias-corrected EVI-estimators are proposed and discussed, together with the use of generalized means in the estimation of parameters of extreme events, like the EVI, a high quantile or even the tail dependence coefficient.The finite-sample behavior as well as robustness regarding sensitivity to data contamination are assessed through Monte-Carlo simulation studies.

Hierarchical Bayes Unit-Level Small Area Estimation Model for Normal Mixture Populations

Shuchi Goyal
University of Georgia
go.shuchi@gmail.com
Coauthors: Gauri Datta, University of Georgia, US Census Bureau; Abhyuday Mandal, University of Georgia

National statistical agencies are regularly required to produce estimates about various subpopulations, formed by demographic and/or geographic classifications, based on a limited number of samples. Traditional direct estimates computed using only sampled data from individual sub-populations are usually unreliable due to small sample sizes. Subpopulations with small samples are termed small areas or small domains. To improve on the less reliable direct estimates, model-based estimates, which borrow information from suitable auxiliary variables, have been extensively proposed in the literature. However, standard model-based estimates rely on the normality assumptions of the error terms. In this research we propose a hierarchical Bayesian (HB) method for the unit-level nested error regression model based on a normal mixture for the unit-level error distribution. Our method proposed here is applicable to model unit-level error outliers and to cases where each small area population is comprised of two subgroups, neither of which can be treated as an outlier. Our proposed method is more robust than the normality based standard HB method (Datta and Ghosh 1991) to handle outliers or multiple subgroups in the population. To implement our proposal we use a uniform prior for the regression parameters, random effects variance parameter, and the mixing proportion, and we use a partially proper non-informative prior distribution for the two unit-level error variance components. We apply our method to predict country areas of corn, originally considered by Battese et al. (1988), and compare these predictions with those obtained by applying the Datta and Ghosh (1991) method and the Chakraborty et al. (2018) method. Our simulation study comparing these three Bayesian methods when the unit-level error distribution is normal, or t, or two-component normal mixture showed the superiority of the proposed method.

Perspectives on Statistical Consulting

Emily Griffith
North Carolina Chapter of the American Statistical Association
eghohmei@ncsu.edu
Coauthors: TBA

This panel discussion will feature several North Carolina statisticians' perspectives on statistical consulting. Panelists will discuss the unique aspects of statistical consulting that they enjoy and share stories from their careers to illustrate their points. The panelists will also offer their opinions on the future of consulting.

Using Mobile Monitors for Fine-Scale Spatiotemporal Air Pollution Analysis

Yawen Guan
NCSU/SAMSI
yawenguan@gmail.com
Coauthors: Maggie Johnson, Matthias Katzfuss, Elizabeth Mannshardt-Hawk,Kyle P Messier, Brian J Reich and Joon Jin Song.

People are increasingly concerned with understanding their personal environment, including possible exposure to harmful air pollutants. In order to make informed decisions on their day-to-day activities, they are interested in real-time information on a localized scale. Thus it is important to understand pollutant patterns on a fine scale, as microenvironmental effects can be extremely varied due to factors such as meteorology and local traffic patterns. A methodological framework utilizing fine-scale measurements to provide real-time air pollution maps as well as short-term air quality forecasts on a fine-resolution spatial scale could prove to be instrumental in increasing public awareness and understanding. The Google Street View study provides a unique source of highly detailed data with spatial and temporal complexities, with the potential

to provide information about commuter exposure and hot spots within city streets with high traffic, as well as complex patterns due to meteorological effects and microenvironments. We develop a computationally-efficient spatiotemporal model for these data and use the model to make high-resolution maps of current air pollution levels and short-term forecasts. We also show via a simulation experiment that mobile networks are far more informative than an equally-sized stationary networks. This modeling framework has important real-world implications in better understanding citizens' personal environments, as data production and real-time availability continue to be driven by the ongoing development and improvement of mobile measurement technologies.

Ensemble Data Assimilation on a Non-Conservative Adaptive Mesh

Colin Guider
UNC Chapel Hill
`cguider1@live.unc.edu`
Coauthors: Ali Aydogdu, Alberto Carrassi, Chris Jones

Solving dynamical systems numerically on an adaptive mesh raises interesting and challenging methodological issues for data assimilation, especially for ensemble-based techniques.

First, the mesh being time-dependent implies that each ensemble member is represented on its own mesh and that, at the analysis times, the values of the physical variables on the mesh must be updated. In addition to the dynamical adaptivity, a re-meshing process can be present, which implies that the mesh dimension is not conserved. These aspects represent fundamental methodological challenges for classical data assimilation methods. Standard approaches for ensemble-based data assimilation rely on a fixed mesh and are thus unsuitable in this case (Guider et al. 2017).

In this work, we first describe the specific challenges for the classical Ensemble Kalman Filter in this context and then present a new method that relies upon the use of a super-mesh that allows for evaluating the ensemble-based error statistic consistently. The new method is applied to a one-dimensional mesh using a low-order model and numerical results are presented.

Weibull-Truncated Exponential Distribution:Properties and Applications

Ahtasham Gul
Department of Statistics, COMSATS University, Lahore, Pakistan
`ahtashamgul@gmail.com`
Coauthors: Muhammad Mohsin, Amir Nadeem

These efforts to modify distributions are attempt to get more pliable models. Weibull and Exponential distributions are widely used in reliability and survival analysis. A new distribution of exponential family has been developed comprising of four parameters named Weibull-Truncated Exponential Distribution (W-TEXPD). Each parameter plays an ample role in modeling the data. A number of distributions are the special cases of proposed distribution. It appears that this model can be used as an substitute to the Exponential, standard Weibull and shifted Weibull distributions. The statistical characteristics namely cumulative distribution function, hazard function, cumulative hazard function, skewness and kurtosis, percentile, entropy and order statistics are discussed. By using maximum likelihood estimation procedure, the unknown parameters of W-TEXPD are computed. The proposed probability distribution is fi

tted to two real data sets to emphasize its application which demonstrates better

fit than Weibull, Gamma, Exponential and truncated exponential distributions.

MSC(2010): 60E05; 62N05. Keywords:Exponential Distribution; Weibull Distribution; W-TEXPD; T-ZT family; hazard function; Shannon entropy; maximum likelihood estimate; order statistics.

Randomized Response Techniques for Estimation of Small Area Total

Sat Gupta
UNC Greensboro
sngupta@uncg.edu
Coauthors: Shakeel Ahmed & Javid Shabbir, Quaid-I-Azam University, Islamabad

In social surveys involving questions that are sensitive or personal in nature, respondents may not provide correct answers to certain questions asked by the interviewer. The impact of this non-response or inaccurate response becomes even more acute in the case of small area estimation where we already have the problem of small sample size coming from the small area. To obtain a truthful response, we use randomized response techniques in each small area. We use the word model in two senses - one in the context of population models, i.e. the relationship between the study variable and the auxiliary variable, and second, the scrambled response model. Application of the scrambled response model increases the variance of the associated estimators in exchange for greater respondent privacy. We focus on the problem of estimating small area total and examine its performance both theoretically and numerically.

Multiresolution Tensor Decomposition for Replicated Spatial Passing Networks

Shaobo Han
Duke University
shaobo.han@duke.edu
Coauthors: David B. Dunson

This research is motivated by soccer positional passing networks collected across multiple games. We refer to these data as replicated spatial passing networks—to accurately model such data it is necessary to take into account the spatial positions of the passer and receiver for each passing event. This spatial registration and replicates that occur across games represent key differences with usual social network data. As a key step before investigating how the passing dynamics influence team performance, we focus on developing methods for summarizing different team's passing strategies. Our proposed approach relies on a novel multiresolution data representation framework and Poisson nonnegative block term decomposition model, which automatically produces coarse-to-fine low-rank network motifs. The proposed methods are applied to detailed passing record data collected from the 2014 FIFA World Cup.

Prenet Penalization in Factor Analysis and its Applications

Kei Hirose
Kyushu University
hirose@imi.kyushu-u.ac.jp
Coauthors: Yoshikazu Terada

We propose a prenet (product elastic net), which is a new penalization method for factor analysis models. The penalty is based on the product of a pair of elements in each row of the loading matrix. The prenet not only shrinks some of the factor loadings toward exactly zero, but also enhances the simplicity of the loading matrix, which plays an important role in the interpretation of the common factors. In particular, with a large amount of prenet penalization, the estimated loading matrix possesses a perfect simple structure, which is known as a desirable structure in terms of the simplicity of the loading matrix. Furthermore, the perfect simple structure estimation via the prenet turns out to be a generalization of the k-means clustering of variables. On the other hand, a mild amount of the penalization approximates a loading matrix estimated by the quartimin rotation, one of the most commonly used oblique rotation techniques. Thus, the proposed penalty bridges a gap between the perfect simple structure and the quartimin rotation. We illustrate the usefulness of our penalty through the analysis of real data.

A Joint Learning of Multiple Precision Matrices with Sign Consistency

Yuan Huang
University of Iowa
yuan-huang@uiowa.edu

The Gaussian graphical model is a popular tool for inferring the relationships among random variables, where the precision matrix has a natural interpretation of conditional independence. With high-dimensional data, sparsity of the precision matrix is often assumed, and various regularization methods have been applied for estimation. Under quite a few important scenarios, it is desirable to conduct the joint estimation of multiple precision matrices. In joint estimation, entries corresponding to the same element of multiple precision matrices form a group, and group regularization methods have been applied for estimation and identification of the sparsity structures. For many practical examples, it can be difficult to interpret the results when parameters within the same group have conflicting signs. To tackle this problem, we develop a regularization method for the joint estimation of multiple precision matrices. It effectively promotes the sign consistency of group parameters and hence can lead to more interpretable results, while still allowing for conflicting signs to achieve full flexibility. Its consistency properties are rigorously established. Simulation shows that the proposed method outperforms the competing alternatives under a variety of settings. With two data example, the proposed method leads to different and more consistent findings.

Generalized Ratio-Type Estimator for Population Variance Using Auxiliary Information In Simple Random Sampling

Muhammad Ismail
COMSATS Institute of Information Technology Lahore, Pakistan
drismail39@gmail.com
Coauthors: Sumbal Zurwa, Nazia Kanwal

This paper suggests a new generalized ratio-type estimator for population variance of study variable utilizing information obtained from two auxiliary variables. Empirically, the estimator proves more efficient than the usual unbiased estimator and the previously existing estimators of Isaki (1983), Upadhyaya and Singh (1999), Kadilar and Cingi (2006), Yadav et al. (2013). Efficiency of the new generalized estimators has been compared mathematically with that of Yadav et al. (2013). The empirical study supports the new estimators against above mentioned estimators.

Big Data Analysis of over 100 million publications using the Microsoft Academic Graph

Darpan Jhawar
Department of Computer Science, UNCG
sdmohant@uncg.edu
Coauthors: Prashanti Manda, Somya D. Mohanty

This study presents a big data analysis of almost 100 million scientific publications from the Microsoft academic graph. We present two distinct yet related analysis here – 1) a novel normalized citation metric that takes into account the age of a publication in addition to the field of study. And 2) a machine learning model that analyzes various characteristics of a publication to predict its "success" as measured by citations.

Normalized Citation Metric: Absolute citation counts of scientific publications can be misleading and cannot be compared across different fields of study primarily because older publications have the advantage of age. In addition, some fields of study tend to have an overall higher citation count while others have lower counts. In order to account for these two primary factors, we developed a Normalized Citation Metric that first normalizes the absolute citation count by age of the publication, and then uses the distribution of citation scores in its field of study to further normalize the impact respective to its field. The resulting normalized scores lie in the range of [0-1] leading to a metric that is consistent across different fields and

ages. We present analytics using this metric that delves trends of scientific impact over time and across different fields.

Prediction of Scientific Impact: We developed a machine learning model based Random Forests that uses various attributes of a publication such as information of the first and last author, affiliations, journal impact, conference venue, etc. to predict the citation impact of the publication. Our results indicate that citation success can be predicted with a 84% accuracy using this model. Some of the top factors for predicting the success of a journal publication include the total number of citations of the journal (across all its publications), the number of publications in the journal, and the rank (impact factor) of the journal. Surprisingly, we found that information pertaining to the authors (such as the authors' total number of citations) was less important as compared to the journal's attributes.

Full-spectrum Copy Number Variation Detection by High-throughput DNA Sequencing

Yuchao Jiang
University of North Carolina, Chapel Hill
`yuchaoj@email.unc.edu`
Coauthors: Rujin Wang, Eugene Urrutia, Ioannis N. Anastopoulos, Katherine L. Nathanson, Nancy R Zhang

Copy number variations (CNV) is an important type of genetic variation that has been associated with diseases. High-throughput DNA sequencing enables detection of CNVs on the genome-wide scale with fine resolution, but suffers from many sources of biases and artifacts that lead to false discoveries and low sensitivity. We describe CODEX2, a statistical framework for full-spectrum CNV profiling that is sensitive for variants with both common and rare population frequencies and that is applicable to study designs with and without negative control samples. CODEX2 can be applied to whole-genome, whole-exome, and targeted sequencing platforms. We demonstrate and evaluate CODEX2 on whole-exome and targeted sequencing data, where biases are the most prominent. On whole-exome sequencing data from the 1000 Genomes Project, CODEX2 outperforms existing methods and, in particular, significantly improves sensitivity for common CNVs. On targeted sequencing data from a case-control study of melanoma patients and cell lines, CODEX2 identifies somatic CNVs in concordance with the results obtained from an independent cohort from the Cancer Genome Atlas. If time permits, our ongoing work on detecting CNVs by single-cell DNA sequencing will also be discussed.

Design Oriented Modeling

Bradley Jones
SAS Institute
`bradley.jones@jmp.com`

Optimal design of experiments (DOE) is about choosing a design that is able to efficiently estimate some specified model (or models). In other words, optimal DOE is model oriented DOE.

This begs the question, why isn't there design oriented modeling? Design oriented modeling explicitly takes the structure of a design in choosing a modeling approach rather than applying some generic algorithm for fitting observational data. This talk will introduce three examples where the structure of a design allows for a modeling approach that takes advantage of the structure of the design.

Semi-Parametric Likelihood Inference for Birnbaum-Saunders Frailty Model.

Kai, Liu
McMaster University
liuk25@math.mcmaster.ca
Coauthors: Balarishnan, N.

Cluster failure time data are commonly encountered in survival analysis due to dif- ferent factors such as shared environmental conditions and genetic similarity. In such cases, careful attention needs to be paid to the correlation among subjects within same clusters. In this talk, we discussed a semi-parametric frailty model based on Birnbaum-Saunders frailty distribution and developed an estimation method using Monte Carlo approximation.

Visualizing IPUMS-I demographics

Adrianna Kallis

amkallis@iastate.edu
Coauthors: Megan Aadland and Silas Bergen

This presentation describes the creation of an interactive visualization dashboard using data from Integrated Public Use Microdata Series International (IPUMS-I). This dashboard will help IPUMS-I users quickly understand trends in population, education, and employment for different countries around the globe. In an attempt to make the data easy to interpret for individuals without an analytics background, we made a visualization displaying educational attainment and employment status by sex and country over time. Additionally, age-sex pyramids allow users to observe changes in population trends over time and across various countries. The visualization makes observing the development status of different countries simple. The visualization is interactive, giving users the power to observe the data in which they are interested. This presentation will describe the entire process behind this project, from cleaning and aggregating the microdata to visualizing the aggregated data.

Ratio Estimation of the Mean Under RRT Models

Geeta Kalucha
P.G.D.A.V. college, University of Delhi
geetakalucha@gmail.com
Coauthors: Qi Zhang , Sadia Khalil

In this paper, we introduce a Geometric Mean ratio estimator for finite population mean in the context of optional RRT models. We discuss the Bias and the Mean Square Error (MSE) of our proposed ratio estimator, correct up to first order of approximation, and present its comparison with some other estimators. A simulation study is also conducted to validate the theoretical results. Both the theoretical and the empirical results show that the proposed ratio estimator is more efficient than the competing estimators.

Locally Optimal Designs for Mixed Continuous and Binary Responses

Kao, Ming-Hung
Arizona State University
mkao3@asu.edu
Coauthors: Kim, Soohyun, and Khoger, Hazar

Experiments with mixed continuous and binary responses are not uncommon in practice. In this work, we are concerned with locally optimal designs for such experiments with a joint model for both continuous and binary responses. A modified complete class approach is proposed for deriving a complete class of locally optimal designs. We also implement an optimization algorithm to search over the complete class for optimal designs. The optimality of the designs that we obtain is verified by the equivalence theorem.

Population Sized Graphical Record Linkage

Andee Kaplan
Duke University
andrea.kaplan@duke.edu
Coauthors: Rebecca C. Steorts

Bayesian graphical record linkage is a method that has been used with great success in the recent past to join messy data from multiple sources in order to facilitate secondary analyses. From in-sample unique entity estimation to logistic regression, these post-linkage analyses benefit from the Bayesian framework through the natural error propagation that is afforded. In this talk, we present an empirically motivated Bayesian graphical record linkage method in conjunction with a Bayesian nonparametric capture-recapture method to provide population size estimation with uncertainty quantification.

"Computationally Viable and Effective Feature Selection in L_0 Norm"

Ana Maria Kenney
Pennsylvania State University
ajk5910@psu.edu
Coauthors: Francesca Chiaromonte, Giovanni Felici

Because of continuous advances in mathematical programing, Mix Integer Optimization has become a competitive vis-a-vis popular regularization method for selecting features in regression problems. The approach exhibits unquestionable foundational appeal and versatility, but also poses important challenges. We tackle these challenges, reducing computational burden when tuning the sparsity bound (a parameter which is critical for effectiveness) and improving performance in the presence of feature collinearity and of signals that vary in nature and strength. Importantly, we render the approach efficient and effective in applications of realistic size and complexity – without resorting to relaxations or heuristics in the optimization, or abandoning rigorous cross-validation tuning. Computational viability and improved performance in subtler scenarios is achieved with a multi-pronged blueprint, leveraging characteristics of the Mixed Integer Programming framework and by means of whitening, a data pre-processing step.

Mean Estimation of Sensitive Variables under Measurement Errors using Optional RRT Models

Sadia Khalil
Department of Statistics, Lahore College for Women University, Lahore, Pakistan
sadia_khalil@hotmail.com
Coauthors: Qi Zhang, Department of Mathematics and Statistics, UNC Greensboro

In this study we propose an improved mean estimator for a sensitive variable under simple random sampling using Optional RRT models when measurement errors are present. We discuss two scrambling options and compare the mean square error (MSE) of the proposed estimator with some of the commonly used estimators. Both theoretical and empirical results show the superiority of the proposed estimator over existing estimators. Models are evaluated both with respect to efficiency and respondent privacy

Efficient Monitoring of Process Mean Using Exponentially Weighted Moving Average Control Charts, Designed With Ratio Estimators under Ranked Set Sampling

Hina Khan
Department of Statistics GC University Lahore, Pakistan
hinakhan@gcu.edu.pk
Coauthors: Saleha Farooq

Focusing on quick detection of small and moderate shifts in the procedural mean, this study proposes EWMA-type control charts by considering some auxiliary information. The ratio estimation technique for the mean with ranked set sampling design is used in designing the control structure of the proposed charts.Here we developed EWMA control charts using some of the popular ratio-type estimators proposed by Cochran (1940), Sisodia & Dwivedi (1981), Singh & Tailor (2003) and Khan et al. (2014), based on ranked set sampling for the process mean to obtain specific ARLs and which suits when small process shifts are of interest. Furthermore the efficiency of these charts for early detection of changes in the processes is also observed. It is observed that among all estimators under study, Khan et al. (2014) estimator detects the shift in mean much earlier. The practical implementation procedure of the proposed control chart is also shown by using a real industrial data set.

An Optimal Systematic Sampling Scheme

Zaheen Khan
Department of Mathematical Sciences, Federal Urdu University of Arts, Science and Technology, Islamabad, Pakistan
zkurdu@gmail.com

This paper focused to gain in efficiency of an estimator of a parameter by means of optimal pairing of units in the systematic sample. This study is more generalized and applicable for all possible choices of population and sample sizes i.e. N and n respectively. Efficiency comparisons with the well-known sampling schemes have also been carried out by theoretically and numerically.

An Optimal Systematic Sampling Scheme

Zaheen Khan
Department of Mathematical Sciences, Federal Urdu University of Arts, Science and Technology, Islamabad, Pakistan.
zkurdu@gmail.com

Abstract This paper focused to gain in efficiency of an estimator of a parameter by means of optimal pairing of units in the systematic sample. This study is more generalized and applicable for all possible choices of population and sample sizes i.e. N and n respectively. Efficiency comparisons with the well-known sampling schemes have also been carried out by theoretically and numerically.

Hurricanes and Rate-Induced Tipping

Claire Kiers
University of North Carolina at Chapel Hill
cekiers@live.unc.edu

In this talk we will analyze Kerry Emanuel's 2-dimensional model of a hurricane, FAST, which describes how surface circular wind speed and inner core moisture evolve over time in a hurricane. We will focus on the effect that two model parameters, wind shear and full potential intensity, have on the strength of a hurricane, and in particular we will look at possibilities of rate-induced tipping in the system if these parameters are allowed to vary over time. Our first main result will show that it is impossible to create a hurricane through rate-induced tipping; some noise must be present in the system for a hurricane to form. On the other hand, rate-induced tipping can cause a hurricane to die, if wind shear and full potential intensity are both increased. We will give an example to show how this can happen.

The Copula Functional ARCH Directional Dependence for Intraday Volatility with High-frequency Financial Data

Jong-Min Kim
University of Minnesota at Morris, USA
jongmink@morris.umn.edu
Coauthors: Sun Young Hwang, Department of Statistics, Sookmyung Women's University, Seoul, South Korea

This paper proposes a copula directional dependence by using a bivariate Gaussian copula beta regression with the Hormann et al. (2013) functional ARCH(1) (fARCH) model to suit high-frequency time series that account for intraday volatilities. With simulated high-frequency data, we show how the copula fARCH directional dependence of intraday volatility can be useful in terms of graphical displays for tick-by-tick price changes in a day. We can perform a test of significance of the copula fARCH directional dependence of intraday volatility by the permutation test, p-value, and bootstrapping confidence interval. To validate our proposed method with real data, we use the Korea Composite Stock Price Index (KOSPI) and the Hyundai-Motor (HD-Motor) company stock data with one minute high-frequency. We show that copula fARCH directional dependence of intraday volatility by B-spline basis function is superior to that by Fourier basis function in terms of the percent relative efficiency of bias and mean squared error. This research shows that the copula functional ARCH directional dependence of intraday volatility can be an important statistical method to illustrate the directional dependence of intraday volatility on the financial market.

De Novo Detection and Accurate Inference of Differentially Methylated Regions

Keegan Korthauer
Dana-Farber Cancer Institute and Harvard T.H. Chan School of Public Health
keegan@jimmy.harvard.edu
Coauthors: Sutirtha Chakraborty, Yuval Benjamini, and Rafael Irizarry

A fundamental task in the analysis of methylation sequencing data is to detect Differentially Methylated Regions (DMRs), a key step in untangling the complex role of epigenetic modification in gene regulation. However, current computational approaches for detecting such differential regions do not provide accurate statistical inference. A major challenge in reporting uncertainty is that a genome-wide scan is involved in detecting regions, which needs to be properly accounted for. A further challenge is that sample sizes are limited due to the cost of the technology. We propose an approach that detects DMRs and assesses their uncertainty in a rigorous manner. Transformed methylation levels are modeled using generalized least squares while accounting for inter-individual and inter-loci variability. Significance of DMRs is assessed against a pooled null distribution that can be implemented even when as few as two samples per population are available. Using both experimental data and Monte Carlo simulation, we show our approach has improved sensitivity to detect regions enriched for downstream changes in gene expression while accurately controlling the False Discovery Rate (FDR). We also highlight the benefits of our inferential approach in the analysis of a recent groundbreaking experiment probing the influence of promoter DNA methylation on transcription.

Persistence Curves: A New Vectorization of Persistence Diagrams

Austin Lawson
UNC - Greensboro
azlawson@uncg.edu
Coauthors: Yu-Min Chung

Topological Data Analysis (TDA) is a field of mathematics concerned with analyzing the shape of data to extract information. The main tool of TDA is Persistent Homology, which is used to generate a summary of the data called a Persistence Diagram. Recently, research in the field has been focused on combining TDA and machine learning through a vectorization of these diagrams. we propose a class of such vectorizations called Persistence Curves. By combining these curves with the support vector machines algorithm, we achieve high accuracy scores on two texture datasets, Outex and UIUCTex.

Using Administrative and Laboratory Data from the Electronic Health Record to Examine Frailty Indicators for Early and Late Readmission among Hospitalized Older Adults

Deborah Lekan, PhD, RN-BC
University of North Carolina at Greensboro, School of Nursing, 409 Moore Nursing Building, Greensboro, NC 27402
dalekan@uncg.edu
Coauthors: Thomas P. McCoy, PhD, PStat, Somya Mohanty, PhD, Prashanti Manda, PhD, Marjorie Jenkins, PhD, RN, NEA-BC, FACHE, & Rohit Gulia, BS

Frailty is a clinical syndrome of impaired homoeostasis and decreased physiologic capacity resulting in diminished ability to resist and recover from physiologic and psychosocial stressors (Lekan & McCoy, 2018; Rodríguez-Mañas & Sinclair, 2014). Hospital readmissions are associated with negative patient outcomes and high health care costs. It is estimated that nearly 20% of Medicare patients are readmitted within 30 days, and 12% of these are potentially avoidable. This retrospective, observational study investigates the use of administrative and laboratory data and mapping of International Statistical Classification of Disease and Related Health Problems (ICD) codes to validate an existing 16-item frailty risk score (FRS) derived from clinical flow sheet data in the electronic health record (EHR) and the additional benefit of variables for dysphagia, physical function, and biological markers including creatinine, sodium, and glucose on accuracy for predicting early (3 and 7-day) and late (30 and 90-day) readmission in over 57,000 adults, 50 years of age and older alive at discharge of their initial hospitalization. Area under the curve (AUC) of receiver operating characteristic (ROC) curves from multivariable logistic regression will be used to quantify accuracy, along with their 95% CIs. Tests of equal AUCs are presented based on methods from DeLong, DeLong, & Clarke-Pearson (1988). Sensitivity analyses with and without weighting based on inverse probability weighting from propensity scores constructed from initial admission characteristics will be examined. Implications for practitioners assessing frailty and risk of early and late readmission will be discussed.

References:

Rodríguez-Mañas, L., & Sinclair, A. J. (2014). Frailty: The quest for new domains, clinical definitions and subtypes. Is this justified on new evidence emerging? Journal of Nutrition Health and Aging, 18(1), 92–94.

Lekan, D. A., & McCoy, T. P. (2018). Frailty risk in hospitalised older adults with and without diabetes mellitus. Journal of Clinical Nursing, (ePub ahead of print). https://doi.org/10.1111/jocn.14529

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics, 44, 837–845.

Simultaneous Inference of Differentially Expressed Isoforms for RNA Sequencing Data

Bo Li

bli@citadel.edu

RNA sequencing technologies provide measurement of the presence and quantity of entire transcriptome of biological samples. In this article, we describe simultaneous inferential methods of detecting differentially expressed gene isoforms based on a Poisson generalized linear model. We derive the joint asymptotic distribution of the pivotal quantity. Since the sample size of RNA sequencing data is small in practice, parametric bootstrap method is outlined and it is shown that it provides improvement for large sample approximation. We demonstrate the validity of the proposed method in detecting differentially expressed isoforms through Monte Carlo simulation. It shows the proposed method controls family wise error rates for large-scale inference. In addition, we compare the empirical power rates under a set of sample sizes. Even though the proposed method can be extended to many experimental designs, we focus on a balanced block design in this article.

Individual Clear Effects for Fractional Factorial Designs

William Li
Shanghai Advanced Institute of Finance
wlli@saif.sjtu.edu.cn
Coauthors: Qi Zhou, Hongquan Xu

In planning a fractional factorial experiment, prior knowledge may suggest that effects involving some factors are more important than others. Literature on how best to incorporate such information when assigning factors to the columns of a design has not received much attention. We propose the concept of individual clear effects (iCE), which is in the same spirit of iWLP of Li et al. (2015) and iGWLP of Li et al. (2018). We discuss how to utilize iCE to assign factors more effectively. Motivated by a real problem, we introduce the clear effects pattern to construct the maximized clear effects pattern (MCEP) designs. These designs are often different from commonly used minimum aberration designs, and many of them also maximize the number of clear two-factor interactions. We then extend the definition of iCE and MCEP designs by considering blocking schemes. Finally, we study some properties of these designs, which can be used to reduce the computational burden for design construction.

Localized Topological Metric for Edge Detection in Images using CNNs

Edgar Lobaton
North Carolina State University
edgar.lobaton@ncsu.edu
Coauthors: Qian Ge

Edge detection is one of the crucial low-level operations in image processing and computer vision. It is commonly used as a pre-processing step for image segmentation, object detection and other high-level computer vision tasks. The performance is often measured by looking at the pixel differences between predicted and groundtruth labels. For segmentation purposes, closed boundaries are desired while small offset of the detection can be tolerated. In this case, it makes sense to ensure that the topological structure of the segmentation is properly captured instead of just measuring pixel differences. We propose a new evaluation metric based on localized persistence homology to measure the difference between two edge maps invariant to small locally bounded deformations. By using this new metric as an additional loss term, we train a Convolutional Neural Network (CNN) to focus on closing edge gaps. To demonstrate the efficiency of the proposed metric, we demonstrate experiments on a synthetic edge detection dataset developed for foraminifera morphological segmentation.

d-QPSO: A Quantum-Behaved Particle Swarm Technique for Finding *D*-Optimal Designs with Discrete and Continuous Factors and a Binary Response

Joshua Lukemire
Emory University
joshlukemire@gmail.com
Coauthors: Abhyuday Mandal and Weng-Kee Wong

Identifying optimal designs for generalized linear models with a binary response can be a challenging task, especially when there are both discrete and continuous independent factors in the model. Theoretical results rarely exist for such models, and for the handful that do, they usually come with restrictive assumptions. In this paper we propose the *d*-QPSO algorithm, a modified version of quantum-behaved particle swarm optimization, to find a variety of *D*-optimal approximate and exact designs for experiments with discrete and continuous factors and a binary response. We show that the *d*-QPSO algorithm can efficiently find locally *D*-optimal designs even for experiments with a large number of factors and robust pseudo-Bayesian designs when nominal values for the model parameters are not available. Additionally, we investigate robustness properties of the *d*-QPSO algorithm-generated designs to various model assumptions and provide real applications to design a bio-plastics odor removal experiment, an electronic static experiment, and a ten-factor car refueling experiment.

Penalized Re-descending M Estimation with Non-Convexity for High-Dimensional Asymmetric Data

Bin Luo
The University of North Carolina at Greensboro
b_luo@uncg.edu
Coauthors: Xiaoli Gao

Penalized M estimation regression including the Huber-Lasso and LAD-Lasso has been widely used when the data is contaminated in y direction. However, when the random errors follow irregular distributions such as asymmetry and heteroscedasticity, in high-dimensional settings, simultaneous mean estimation and variable selection are still of interest in many applications. In this paper, we study the high-dimensional mean regression using non-convex penalized re-descending M estimators, denoted as PRAM estimators, in general irregular settings such as the random errors are lack of symmetry and homogeneity, and the regressors are lack of sub-Gaussian assumption. To reduce the bias caused by the irregular random errors, the PRAM

estimators use a family of loss function with strong robustness and diverging parameters to approximate the mean function from the traditional quadratic loss. The PRAM estimators are investigated in this paper both theoretically and numerically. In theory, we show that, in the ultra-high dimension setting where the dimensionality can grow exponentially with the sample size, with certain mild conditions on the covariate and loss functions with diverging parameters, the PRAM estimators have local estimation consistency at the minimax rate enjoyed by LS-Lasso. In simulation, we numerically demonstrate the performance of the PRAM estimators using three types of loss functions all equipped with diverging parameters (the Huber loss, Tukey's biweight loss and Cauchy loss) and two types of penalty functions (the Lasso and MCP penalties). Our simulation studies exhibit satisfactory finite sample performance of the PRAM estimators under general irregular settings.

A Fused Gaussian Process Model for Very Large Spatial Data

Pulong Ma
The Statistical and Applied Mathematical Sciences Institue and Duke University
pma@samsi.info
Coauthors: Emily L. Kang, University of Cincinnati

With the development of new remote sensing technology, large or even massive spatial datasets covering the globe become available. Statistical analysis of such data is challenging. This article proposes a semiparametric approach to model large or massive spatial datasets. In particular, a Gaussian process with additive components is proposed, with its covariance structure consisting of two components: one component is flexible without assuming a specific parametric covariance function but is able to achieve dimension reduction; the other is parametric and simultaneously induces sparsity. The inference algorithm for parameter estimation and spatial prediction is devised. The resulting spatial prediction method that we call fused Gaussian process (FGP), is applied to simulated data and a massive satellite dataset. The results demonstrate the computational and inferential benefits of the FGP over competing methods and show that the FGP is more flexible and robust against model misspecification.

A Modified Ratio Estimator of Population Mean of a Sensitive Variable in the Presence of Non-Response in Simple Random Sampling

Mahnaz Makhdum
Department of Statistics, Lahore College for Women University, Lahore, Pakistan
minz_mak@hotmail.com
Coauthors: Aamir Sanaullah, COMSATS Institute of Information and Technology, Lahore, Pakistan; Muhammad Hanif, National College of Business Administration and Economics, Lahore, Pakistan

A modified ratio estimator has been proposed for finite population mean of a sensitive study variable in the presence of non-response under a Simple Random Sampling Without Replacement Design (SRSWOR) using the Randomized Response Technique (RRT) in two phase sampling . Auxiliary information from a positively correlated non-sensitive variable is used. The expressions for Bias and MSE of the proposed estimator are derived. The performance of the proposed estimator is evaluated both theoretically and empirically.

EzGP: Easy-to-Interpret Gaussian Process Models for Computer Experiments with Both Quantitative and Qualitative Factors

Abhyuday Mandal
University of Georgia
amandal@stat.uga.edu
Coauthors: Qian Xiao, Chunfang Devon Lin and Xinwei Deng

Computer experiments with both quantitative and qualitative inputs are commonly used in science and engineering applications. Constructing desirable emulators for such computer experiments remains a challenging problem. Here we propose an easy-to-interpret Gaussian process (EzGP) model for computer experiments to reflect the change of the computer model under different level combinations of qualitative factors. The proposed modeling strategy, based on an additive Gaussian process, is flexible to address the heterogeneity of computer models involving qualitative factors. In addition, we develop two useful variants of the EzGP model to achieve computation efficiency and to deal with large data size. The merits of these models are illustrated by a real data application and several numerical examples.

Distributions of Persistence Diagrams and Approximations

Vasileios Maroulas
University of Tennessee
vmaroula@utk.edu
Coauthors: Josh Mike

In this talk, a nonparametric way is introduced to estimate the global probability density function of a random persistence diagram. A kernel density function centered at a given persistence diagram and a given bandwidth is constructed. Our approach encapsulates the number of topological features and considers the appearance or disappearance of features near the diagonal in a stable fashion. In particular, the structure of our kernel individually tracks long persistence features, while considering features near the diagonal as a collective unit. The choice to describe short persistence features as a group reduces computation time while simultaneously retaining accuracy. Indeed, we prove that the associated kernel density estimate converges to the true distribution as the number of persistence diagrams increases and the bandwidth shrinks accordingly. We also establish the convergence of the mean absolute deviation estimate, defined according to the bottleneck metric. Lastly, examples of kernel density estimation are presented for typical underlying datasets.

Detecting Variability in Genome with Applications to Public Health

Sunil Mathur
Texas A&M University-Corpus Christi
Sunil.Mathur@tamucc.edu

Since the progress made in human genome, genomics has emerged as very challenging and promising field. The human genome is one of the many complete genome sequences known which gives us information about the differences from each other, similarities with each other, influence of environment and many other things. Genomics offers us to some extent avoid having some diseases, postpone progress of diseases and eradicate some genetic diseases. Our understanding of genomics and its relevance to medicine may help us to develop targeted drug therapies. The detection of differences in human genome under two different conditions is very important in genomics. The impact of genes and their interaction with behavior, diet, and the environment on the population's health creates a potential source of variance in genomic experiments. We propose a statistical test to detect the variability without assuming the normality condition. The detection of variability in human genome will allow us to identify experimental variables such as behavior, diet, and the environment that affect different biological processes and accuracy of genomic measurements. This research will help in assess the interrelationships among genes, the environment, behavior, and population health.

An Efficient Multiplicity Adjustment for Large Scale Chi-Square Endpoints

Melinda McCann
Oklahoma State University
mccann@okstate.edu
Coauthors: Amy Wagler, University of Texas El Paso

Whenever overall conclusions are warranted, the familywise error rate of a set of inferences requires control. Many methods exist for controlling multiplicity for independent or normally distributed endpoints, but relatively few address non-normal correlated endpoints. This manuscript proposed a fast an simple multiplicity adjustment that strictly controls the type I error for a family of chi-square distributed endpoints. The method is flexible and may be efficiently applied to large throughput chi-square distributed endpoints with any correlation structure. Numerical results confirm that this procedure is effective at controlling familywise error, is far more efficient than utilizing a Bonferroni adjustment, is more computationally feasible in high-dimensional settings than existing methods, and, except for highly correlated data, performs similarly to simulation-based methods. An application illustrates the use of the proposed multiplicity adjustment to a large scale testing example.

Utilizing the Block Diagonal Covariance Structure of Nonregular Two-Level Designs

Robert Mee
University of Tennessee and Nankai University
rmee@utk.edu
Coauthors: David Edwards

Two-level fractional factorial designs are often used in screening scenarios to identify active factors. This presentation shows how certain nonregular two-level designs have information matrices with a block diagonal structure. This structure is appealing since estimates of parameters belonging to different diagonal submatrices are uncorrelated. As such, the covariance matrix of the least squares estimates is simplified and the number of linear dependencies is reduced. The block structure depends on the number of parallel flats, which is easily determined through the indicator function. Nonregular fractional factorial designs can be constructed by combining regular fractions - and these may be performed sequentially or as a single design. Alternatively, nonregular designs can be constructed using generators consisting of linear combinations of interactions. Both means of construction are explored.

Robust Trend Filtering and Outlier Detection

Austin Miller
University of Wyoming & UNCG Statistics REU
amille81@uwyo.edu
Coauthors: Philip Zhu, and Xiaoli Gao

Current linear trend filtering techniques are sensitive to outliers and have no ability to produce robust linear trend estimations. In this manuscript, we propose a method of linear trend filtering that can simultaneously produce robust trend estimates and run sequential outlier detection. We accomplish this goal through the integration of a penalized weighted least squares (Gao & Fang 2017) approach applied to current linear trend filtering techniques. By introducing a second lasso type penalty to L1 trend filtering (Kim et al. 2009) we are able to simultaneously control the number of change points identified in a trend and control the number of outliers identified. We evaluate our method's ability to provide robust trend estimates using both numerical simulations and real data analysis. Our real data analysis is done on the historical real price of the S&P 500. Finding unexplained outliers in stock prices may help to identify market manipulations like insider trading.

A Data-Driven Analysis of Patient Rehospitalization Risk - Machine Learning with Electronic Health Records

Somya D. Mohanty
University of North Carolina - Greensboro
sdmohant@uncg.edu
Coauthors: Rohit Gulia, Deborah Lekan, Prashanti Manda, Thomas McCoy

The current state of art in healthcare and nursing considers a limited set of factors contributing to rehospitalization of patients. Electronic healthcare records provide a rich portfolio of patient information that encompass physical, demographic, and clinical measurements, along with past history of risk and disease. This repository of data can be analyzed effectively to understand and identify various risk factors that predict rehospitalization. Our study explores a data-driven approach towards developing machine learning models capable of identifying at-risk patients for rehospitalization. Utilizing Electronic Health Records (EHR) of over 73,000 patients, containing demographic, laboratory, procedural, and diagnosis (ICD) variables, we evaluate supervised machine learning models towards predicting, 3, 7, 15, and 30 day readmission for patients. The study also compares multiple mdaodels based on parametric (Logistic Regression) and non-parametric (Random Forest and Gradient Boosted Tree) approaches on their performance metrics (AUC and F1-scores), and helps identify key variables which play an important role in prediction of readmission. The goal of the study is to provide hospital administrators and medical practitioners with an automated diagnosis augmentation tool capable of predicting vulnerable patients using a data-driven methodology.

Computing Happiness from Textual Data

Sayed Mostafa
Mathematics Department, North Carolina A&T State University, Greensboro, NC, USA
sayed.mostafa@okstate.edu
Coauthors: Emad Mohamed; University of Wolverhampton, Wolverhampton, UK

We use a corpus of about 100,000 happy moments written by people of different genders, marital statuses, parenthood statuses, and ages to explore the following questions: (1) Can gender/marital status/parenthood and age be predicted from textual data? And (2) are there differences between men and women, married people and non-married people, parents and non-parents and different age groups in terms of their causes of happiness? For the prediction task, we use character, lexical, grammatical, semantic, and syntactic features in a machine learning document classification approach. The explanation of each response variable (gender/parenthood/ marital status/age group) is done using multi-variable binary/multinomial logistic regression to rank a set of predictors in terms of their influence on each specific response variable. This research aims to bring together elements from philosophy and psychology to be examined by computational corpus linguistics methods in a way that promotes the use of Natural Language Processing for the Humanities.

Finite Population Model-Assisted Estimation using Combined Parametric and Nonparametric Regression Smoothers

Sayed Mostafa
Mathematics Department, North Carolina A&T State University, Greensboro, NC, USA
sayed.mostafa@okstate.edu
Coauthors: Qingsong Shan; School of Statistics, Jiangxi University of Finance and Economics, China

This paper considers estimating finite population totals from complex sample surveys in the presence of auxiliary information. Model-assisted estimators which assume a working regression model relating the study variable with the auxiliary data, are common in this context. Both parametric and nonparametric working models have been utilized individually in constructing several model-assisted estimators. Model-assisted estimators with parametric working models are known to be efficient when the assumed working model is correctly specified while using nonparametric smoothers gives more robust estimates but requires relatively large sample sizes. In this paper, we consider the situation where the researcher has an idea of which parametric model can describe the relationship between the study variable and the auxiliary data, but this model may not be adequate in some areas of the data range. Using combined parametric and nonparametric regression smoothers for the working model, we introduce a new class of model-assisted estimators for finite population totals. The proposed estimators are shown to have the desirable asymptotic properties of traditional model-assisted estimators of population totals. The finite sample performance of the new estimators is studied via an empirical study where both simulated and real populations are considered. The empirical results suggest that our proposed estimators are superior to other model-assisted estimators as well as the customary Horvitz-Thompson estimator especially when our knowledge about the form of the working model is reasonable but not necessarily perfect.

How Many Directions Determine a Shape and other Sufficiency Results for Two Topological Transforms

Shayan Mukherjee
Duke University
`sayan@stat.duke.edu`
Coauthors: Justin Curry, Katharine Turner

In this paper we consider two topological transforms based on Euler calculus: the persistent homology transform (PHT) and the Euler characteristic transform (ECT). Both of these transforms are of interest for their mathematical properties as well as their applications to science and engineering, because they provide a way of summarizing shapes in a topological, yet quantitative, way. Both transforms take a shape, viewed as a tame subset M of R^d, and associates to each direction v in S^{d-1} a shape summary obtained by scanning M in the direction v. These shape summaries are either persistence diagrams or piecewise constant integer valued functions called Euler curves. By using an inversion theorem of Schapira, we show that both transforms are injective on the space of shapes – each shape has a unique transform. By making use of a stratified space structure on the sphere, induced by hyperplane divisions, we prove additional uniqueness results in terms of distributions on the space of Euler curves. Finally, our main result provides the first (to our knowledge) finite bound required to specify any shape in certain uncountable families of shapes, bounded below by curvature. This result is perhaps best appreciated in terms of shattering number or the perspective that any point in these particular moduli spaces of shapes is indexed using a tree of finite depth.

The Negative Impact of Outliers in Linear Regression Model; Possible Solutions Using Robust Regression

Mubbasher Munir
Department of Quantitative Methods, University of Magaement and Technology Lahore Pakistan
`mubbasher.munir@umt.edu.pk`

Simple linear regression is unique method and gives affected results when data follows normality. The Bernoulli in 1777 was first one who discussed about the outlier and suggested the removal of the outlying observation from data (Beckman & Cook, 1983). Outlier(s) makes serious effects in statistical modeling and inferences. Even one outlying observation can destroy least squares estimation, resulting in parameter estimates that do not provide useful information for the majority of the data. Robust regression analyses have been developed as an improvement to least squares estimation in the presence of outliers. We have discussed many robust regression estimators like M estimator by Huber (1981), Bi-square, Hampel (1986) and MM estimator (Yohai, 1987) are being used for this research. Furthermore LMS estimators (Rousseeuw, 1984), LAD estimator (Hao & Naiman, 2007), LTS estimators (Rousseeuw, 1984) and S estimators (Hampel, 1975) were also used to compare the more efficient estimator in the presence of outliers. The results shown that the influential points were found major cause of disturbance in the OLS regression estimation but the more precise results were found by using M-estimators (Bi-square) in the presence of outlier(s). The effects of increasing sample size in regression estimation were also checked for more efficient results. Key Words:

Robust Regression, Ordinary Least Square, Breakdown Point, Outlier, M estimator (Huber, Hampel and Bi-square)and Influential Points.

Nonparametric Sub-Sampling for Big Data

Abigael Nachtsheim
Arizona State University
anachtsh@asu.edu
Coauthors: John Stufken

The desire to build predictive models based on datasets with millions or tens of millions of observations is not uncommon today. However, with large datasets, standard statistical methods for analysis and model building can become infeasible due to computational limitations. A standard approach is to obtain a random sample of suitable size and build a predictive model based on the sample. Alternatively, model-based sampling approaches can be employed, but these methods are dependent on a priori knowledge of the correct form of the model for the response of interest (Wang et al., 2017). However, such assumptions are frequently inapplicable, particularly in the big data context. In this paper we explore two new methods of subdata selection that do not require model assumptions. These two proposed approaches use k-means clustering and space-filling designs in an attempt to spread the subdata uniformly throughout the region of the full data. We perform a simulation study and an analysis of real data to investigate the efficacy of the predictive models that result from these methods.

Construction, Properties, and Analysis of Supersaturated Designs Based on Kronecker Products

Christopher Nachtsheim
University of Minnesota
nacht001@umn.edu
Coauthors: Bradley Jones, Ryan Lekivetz, Dibyen Majumdar, Jon Stallings

In this paper, we propose a new method for constructing supersaturated designs that is based on the Kronecker product of two matrices. The motivation for these designs was originally to create a supersaturated design having a few (e.g., three to five) columns that are orthogonal to the others that would not be assigned to any factors, but could remain unlabeled and be used to provide an unbiased estimate of the variance. The approach that we have developed is more general. The construction method leads to a partitioning of the columns of the design such that the columns within a group are correlated to the others within the same group, but are orthogonal to any factor in any other group. We leverage this structure in order to develop an effective model selection procedure. We show that a variant of this procedure can be particularly effective in group screening: unlike previous group-screening procedures, with our designs, main effects in a group are not completely confounded.

Bayesian Zero-Inflated Negative Binomial Regression Based on Pólya-Gamma Mixtures

Brian Neelon
Medical University of South Carolina
neelon@musc.edu

Motivated by a study examining spatiotemporal patterns in inpatient length of stay, we propose an efficient Bayesian approach for fitting zero-inflated negative binomial models. To facilitate posterior sampling, we introduce a set of latent variables that are represented as scale mixtures of normals, where the precision terms follow independent Pólya-Gamma distributions. Conditional on the latent variables, inference proceeds via straightforward Gibbs sampling. For fixed-effects models, our approach is comparable to existing methods. However, our model can accommodate more complex data structures, including multivariate and spatiotemporal data, settings in which current approaches often fail due to computational challenges. Using simulation studies, we highlight key features of the method and compare its performance to other estimation procedures. We apply the approach to a spatiotemporal analysis examining the duration of inpatient stays among United States veterans with type 2 diabetes.

Assessing the Descriptive Epidemiology of Idiopathic Clubfoot in Iowa

Siri Neerchal
University of Maryland College Park
siri@terpmail.umd.edu

Idiopathic talipes equinovarus, or clubfoot, is a common musculoskeletal birth defect. Despite nearly 50 years of study, major risk factors for clubfoot (other than perhaps cigarette smoking during pregnancy) remain elusive. The goals of this project were to examine trends in prevalence, identify geographic hotspots, and estimate associations with selected child and parental characteristics for clubfoot using Iowa Registry for Congenital and Inherited Disorders (IRCID) clubfoot data and Iowa birth data from 1997-2016. The IRCID conducts statewide surveillance for major structural birth defects diagnosed among pregnancies of Iowa residents. A simple linear regression model and a spline model were constructed to estimate statewide prevalence over the 20-year time period, and multilevel Poisson regression analyses were used to estimate relative risk of clubfoot over space and time. Prevalence ratios for various child and parental characteristics were estimated using logistic regression analyses. Sub-analyses stratified by clubfoot laterality were also performed.

It's Not What We Said, It's Not What They Heard, It's What They Say They Heard

Barry D. Nussbaum
American Statistical Association
StatisticsBarry@gmail.com

Statisticians have long known that success in our profession frequently depends on our ability to succinctly explain our results so decision makers may correctly integrate our efforts into their actions. However, this is no longer enough. While we still must make sure that we carefully present results and conclusions, the real difficulty is what the recipient thinks we just said. This presentation will discuss what to do, and what not to do. Examples, including those used in court cases, executive documents, and material presented for the President of the United States will illustrate the principles.

Destructive Cure Rate Model Based On Multiple Treatments

Suvra Pal
University of Texas at Arlington
suvra.pal@uta.edu
Coauthors: N. Balakrishnan and F. Milienos

In this talk, I will first discuss about the cure rate model and then extend the model to the so-called destructive cure rate model, which assumes that each competing cause goes through a destructive process after an initial treatment. Next, I will generalize the destructive cure rate model by assuming that each competing cause undergoes a destructive process for more than one time. For this generalized model, I will present the likelihood inference through the expectation maximization algorithm. Finally, I will present some simulation study results to demonstrate the performance of the model and the estimation procedure.

Assessing Health Disparities using Nonparametric Multivariate Density Estimation subject to Marginal Unimodality Constraints

Rajib Paul
Department of Public Health Sciences, University of North Carolina at Charlotte
Rajib.Paul@uncc.edu
Coauthors: Sujit K. Ghosh (North Carolina State University) and Amy B. Curtis (Hawaii State Department of Health)

Empirical analyses of observed pairs of birthweight and gestational age suggest that underlying joint distribution of these variables are highly dependent and each one of them is marginally unimodal and left skewed. Standard linear regression models are not appropriate for these data. Marginal parametric transformations are quite popular, however, they often lack interpretability and flexibility. We develop a nonparametric density estimation method for assessing the effects of contributing socioeconomic factors tied to risk of low birthweight and preterm birth. The joint density of the bivariate outcome variables is estimated using a flexible class of mixtures of scaled Beta distributions where the mixing weights are suitably constrained to ensure marginal unimodality. Large Sample Consistency of the proposed density estimator is established using the method of sieves and Argmax Continuous Mapping Theorem. The proposed method enlarges the scope of previous studies in two important aspects: (i) joint density provides simultaneous estimation of higher order moments like mean and variance functions; and (ii) marginal shape constraints (unimidality and left skewness) provide more efficient estimates. The proposed procedures are illustrated using simulated and real case study datasets which are shown to successfully identify and characterize the socioeconomic disparities present in gestational age and infant birthweight.

On the probability distribution of durations of heatwaves

Sohini Raha
North Carolina State University
sraha@ncsu.edu
Coauthors: Sujit Ghosh

Characterization of heatwaves is becoming increasingly important in environmental research as they pose a significant threat to many human lives worldwide. Though several quantifications of the extremities of a heatwave have been proposed in literature, they are mostly improvised and there does not exist a universally accepted definition of heatwave. In this paper, we devise a probabilistic inferential framework to characterize heatwave, and come up with a definition which can capture the essence of all existing ad hoc definitions. Based on results for sums of dependent Bernoulli random variables, we derive an approximate distribution on the frequency of such durations for a stationary time series. We select a daily time series e.g. maximum ambient temperature or heat-index (based on temperature and relative humidity) and define "Duration" as the amount of days the time series stays above a chosen threshold in one up-crossing in a fixed location. We

then propose a hierarchical model for the durations and validate it using two different datasets, one with Atlanta data, and the other one with 126 USCRN weather stations spread across the United States. Using the distributions of the durations, we compute the expected duration of an up-crossing corresponding to a threshold in a fixed location, and define an up-crossing to be a heatwave if the duration of that exceeds the expectation. Moreover, we demonstrate a quadratic relationship between the threshold quantiles and the expected duration which makes it easier to identify the heatwaves at any given level of the quantiles of the time series that are generally used to define extreme heatwave.

Modeling Between-Study Heterogeneity for Improved Reproducibility in Gene Signature Selection and Clinical Prediction

Naim Rashid
UNC-CH Biostatistics
naim@unc.edu
Coauthors: Quefeng Li, Jen Jen Yeh, Joseph Ibrahim

In the genomic era, the identification of gene signatures associated with disease is of significant interest. Such signatures are often used to predict clinical outcomes in new patients and aid clinical decision-making. However, recent studies have shown that gene signatures are often not reproducible. This occurrence has practical implications in the generalizability and clinical applicability of such signatures. To improve reproducibility, we introduce a novel approach to select gene signatures from multiple data sets whose effects are consistently non-zero by accounting for between-study heterogeneity. We build our model upon some robust platform-independent quantities, enabling integration over different platforms of genomic data. A high dimensional penalized Generalized Linear Mixed Model (pGLMM) is used to select gene signa- tures and address data heterogeneity. We compare our method to two commonly used strategies that select gene signatures ignoring between-study heterogeneity, and show that these strategies have inferior performance in predicting outcome in new studies. We provide asymptotic results justifying the performance of our method and demonstrate its advantage through thorough simulation studies. Lastly, we motivate our method through a case study subtyping pancreatic cancer patients from four studies using different gene expression platforms.

Markov Chains, Mixing Times, and Couplings

James Marshall Reber
Purdue University
marsha71@purdue.edu

Markov chain mixing times have been of great interest in recent times. A classic mixing time result, due to Diaconis, is on how many riffle shuffles are required to shuffle a deck of n cards - the required number is $3/2 \cdot \log_2(n)$. In this talk, we explore the mixing times of random walks on various graphs using a combinatorial method called coupling. In particular, we give upper bounds on the mixing times of simple random walks on certain families of three-regular graphs, and conjecture some possible generalizations.

Simultaneous Confidence Intervals for Comparing Scale Parameters using Deviances

Scott J Richter
University of North Carolina at Greensboro
sjricht2@uncg.edu
Coauthors: Melinda H. McCann, Oklahoma State University

Permutation confidence intervals based on medians are examined for pairwise comparison of scale. Methods that have been found in the literature to be effective for comparing scale for two groups are extended to the case of all pairwise comparisons, using the Tukey-type adjustment of Richter & McCann (2007) to guarantee strong Type I error rate control. Power and Type I error rate estimates are computed using simulated data. A method based on the ratio of deviances appears to perform best overall.

Tails of distributions – classification and testing

Javier Rojo
Oregon State University
javier.rojo@oregonstate.edu

This introductory lecture provides a systematic presentation of the various classification approaches for probability distributions in terms of their tail heaviness. The concept of tail-heaviness has generated a lot of interest throughout the years. A google search can generate more than 16 million documents for the query "long tail statistics". The lecture will make connections among some of the classification methods and conclude that the method based on the residual life function provides the clearer and more convenient method of classification. Additionally, a test for medium tails vs heavy tails is proposed and some of its operating characteristics are discussed.

Efficient Generation of Unlabeled Graphs

James Rudzinski
UNC Greensboro
jerudzin@uncg.edu

We will discuss two algorithms of McKay related to graph isomorphism. The first algorithm, known as nauty, is widely used to find canonical graph isomorphs along with graph automorphisms. The second is an algorithm for generating a complete set of unique representatives of isomorphism classes of objects. We examine the second algorithm in particular with respect to graph generation and will discuss an efficient variation of this algorithm that can generate distinct representatives for each unlabeled graph on a fixed number of vertices.

An Algebra for the Conditional Main Effects Parameterization

Arman Sabbaghi
Purdue University Department of Statistics
sabbaghi@purdue.edu

The conditional main effect (CME) parameterization system resolves the long-standing aliasing dilemma of the traditional orthogonal components system for two-level regular fractional factorial designs. However, the algebra of the CME system is not yet fully understood, which impedes the development of general results on this system that possess a broad scope of application across designs. We establish a comprehensive algebra for the CME system based on indicator functions. Our algebra facilitates the derivations of general partial aliasing relations for a wide variety of two-level designs. By means of our algebra, we illuminate the implications of traditional design criteria under the CME system for resolution IV designs. A novel feature of our algebra is that it enables immediate and simple D-efficiency calculations for two-level regular designs and models consisting of multiple conditional and traditional effects.

Efficient Numerical Algorithms for the Generalized Langevin Equation

Matthias Sachs
Duke University / SAMSI
msachs@math.duke.edu
Coauthors: Benedict Leimkuhler, University of Edinburgh

We discuss the design and implementation of numerical methods to solve the generalized Langevin equation (GLE) focusing on canonical sampling properties of numerical integrators. For this purpose, we cast the GLE in an extended phase space formulation and derive a family of splitting methods which generalize existing Langevin dynamics integration methods. We show that the dynamics of a suggested integration scheme is consistent with asymptotic limits of the exact dynamics and can reproduce (in the short memory limit) a superconvergence property for the analogous splitting of Langevin dynamics. We apply our proposed integration method to several model systems, including a simple Bayesian inference problem. Using a parameterization of the memory kernel in the GLE as proposed by Ceriotti et al [1], we find that our proposed integration scheme outperforms other previously proposed GLE integration schemes in terms of the accuracy of sampling. Moreover, our experiments indicate the potential benefits of a GLE-based method in comparison to other white noise Langevin dynamics integration schemes in terms of robustness and efficiency.

[1] M. Ceriotti, G. Bussi, and M. Parrinello, "Colored-noise thermostats àla Carte," J. Chem. Theory Comput., vol. 6, no. 4, pp. 1170–1180, 2010.

Statistical Topography for Sea Ice Modeling

Christian Sampson
SAMSI
christian.sampson@gmail.com

Sea ice is an important component of the Earth's climate system whose large scale- long time behavior can depend on smaller scale -short time physical processes which are difficult to directly model. One example is meltponds, ponds which form atop Arctic sea ice from snow melt in the summer months. Meltponds exhibit complex geometries which affect ice albedo and ice strength, two critical parameters for sea ice models. Meltponds also cause error in passive microwave satellite derived summer time sea ice concentrations. This is due to the face that the ponds have same microwave signature as open water. These concentration values make up a 30 year record of ice data and are often used for model calibration. Simple methods of generating melt pond geometries are thus of interest both to parameterize difficult model processes and to find ways to improve summer time sea ice concentration retrievals. We investigate some stochastic methods for generating realistic meltpond geometries suitable for use in sea ice modeling.

Variable Selection for Deterministic Computer Simulator Output

Thomas Santner
The Ohio State University
santner.1@osu.edu
Coauthors: Casey Davis and Christopher Hans

This talk will describe variable selection methodology to identify the "active" inputs of a deterministic simulator code. While the basic methodology can be used in many settings, the description here will use a Bayesian composite Gaussian Process (GP) model. The model assumes that the simulator output can be described as the sum of draw from a GP that incorporates the long-range mean of the output plus a draw from an independent GP that describes small-scale deviations from the mean. A prior is placed on the model parameters that insures the process describing the mean is smoother than that describing local deviations from the mean. Based on a Bayesian fit to this model in which the correlations are assumed to have a Gaussian correlation function, the inputs having smaller estimated correlation parameters are judged to be more active. A reference inactive input is added to the data to judge the size of the correlation parameter for inactive inputs. The importance of design is indicated.

Comparison of Machine Learning Algorithms for Rapid Evaporative Ionization Mass Spectrometry (REIMS) Studies

Amelia Schroeder
East Tennessee State University
schroedera@etsu.edu
Coauthors: Dr. Jessica Prenni, Dr. Julia Sharp, Devin Gredell, and Soo-Young Kim

Traditional methods of beef quality assessment are subjective, inconsistent with consumer preference, and primarily based on gross phenotypic differences resulting in the need for new methods of food authenticity testing and sample identification. In this study, we investigate machine learning algorithms for the prediction of beef quality attributes using the molecular profile of each sample to determine the optimal modeling approach. Rapid Evaporative Ionization Mass Spectrometry (REIMS) is an emerging technique that was used to collect the molecular profile of each sample. Machine learning algorithms were applied to the REIMS data to generate predictive models for the classification of beef quality attributes (i.e. meat grade, tenderness class, angus phenotype). Due to the high volume of predictors within each profile, various dimension reduction techniques were also explored. The results obtained in this study show that high prediction accuracy can be achieved for the classification of beef samples with the use of machine learning and REIMS analysis.

Use of Successive Sampling Strategy for Finite Population Distribution Function Under Nonresponse

Javid Shabbir
Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan
js@qau.edu.pk
Coauthors: Sat Gupta, Department of Mathematics and Statistics, University of North Carolina at Greensboro, USA

Many surveys are being repeated at different time periods for estimating the same characteristic of interest. Generally, values of the study variable change over time and also the values recorded for a given occasion do not provide necessary information at different occasions. Therefore, we address the problem of finite population distribution function at two successive occasions in the presence of non-response. A cost function is used to get an idea about the survey cost in the context of successive sampling on two occasions.

atsnp Search: a Web Resource for Statistically Evaluating Influence of Human Genetic Variation on Transcription Factor Binding

Sunyoung Shin
University of Texas at Dallas
sunyoung.shin@utdallas.edu
Coauthors: Rebecca Hudson, Christopher Harrison, Mark Craven, and Sunduz Keles

Understanding the regulatory roles of non-coding genetic variants has become a central goal for interpreting results of genome-wide association studies. The regulatory significance of the variants may be interrogated by assessing their influence on transcription factor binding. We have developed atSNP Search, a comprehensive web database for evaluating motif matches to the human genome with both reference and variant alleles and assessing the overall significance of the variant alterations on the motif matches. Convenient search features, comprehensive search outputs, and a useful help menu are key components of atSNP Search. atSNP Search enables convenient interpretation of regulatory variants by statistical significance testing and composite logo plots, which are graphical representations of motif matches with the reference and variant alleles. Existing motif-based regulatory variant discovery tools only consider a limited pool of variants due to storage or other limitations. In contrast, atSNP Search users can test more than 37 billion variant-motif pairs with marginal significance in motif matches or match alteration. Computational evidence from atSNP Search, when combined with experimental validation, may help with the discovery of underlying disease mechanisms.

Improved Estimation Procedures of Population Parameter for Sensitive Characteristic using Randomized Response Technique

G N Singh
Department of Applied Mathematics, IIT(ISM), Dhanbad-826004, India
gnsingh_ism@yahoo.com
Coauthors: Surbhi Suman

The present work deals with the problem of estimation of population parameter using randomized response technique when non-response or misrepresentative response occurs during sample surveys due to sensitive nature of attribute (characteristic) under study. The properties of the resultant estimators have been deeply examined. The measure of privacy protection of respondent is also discussed under the suggested estimation procedure. The empirical studies have been performed to demonstrate the efficacy of the proposed estimator over contemporary existing estimators. Suitable recommendations have been made for survey practitioners.

Pseudo Generalized Youden Designs

Rakhi Singh
IITB-Monash Research Academy
agrakhi@gmail.com
Coauthors: Ashish Das; Daniel Horsley

Youden square designs, or Youden rectangles, are classical objects in design theory. Extensions of these were introduced in 1958 by Kiefer and in 1981 by Cheng, in the form of generalized Youden designs (GYDs) and pseudo Youden designs (PYDs), respectively. In this talk I will introduce a common generalization of both these objects, which we call a pseudo generalized Youden design (PGYD). PGYDs share the statistically-desirable optimality properties of GYDs and PYDs, and we show that they exist in situations where neither GYDs nor PYDs do. We determine some numerical necessary conditions for the existence of PGYDs, classify their existence for small parameter sets, and provide constructions for families of PGYDs using patchwork methods based on affine planes.

A Data-Driven Approach to Combinatorial Game Theory

Michael A Smith
Department of Mathematics, Purdue University
smit2589@purdue.edu
Coauthors: Bret Benesh, Department of Mathematics, The College of St. Benedict/St. John's University; Jamylle Carter, Department of Mathematics, Diablo Valley College; Deidra A. Coleman, Department of Mathematics, Wofford College; Douglas G. Crabill, Department of Statistics, Purdue University; Jack H. Good, Department of Computer Science (undergraduate student), Purdue University; Jennifer Travis, Department of Mathematics, Lone Star College; Mark Daniel Ward, Department of Statistics, Purdue University

In a two-player subtraction game, players remove stones from a pile until one player is stuck. The other player is the winner. Subtraction sets of allowable moves characterize each game. Sets of size one and two have been understood since the 1970's. However, sets of size three, S=(x,y,z), do not have precise characterizations. We have utilized 37 years of computation and more than 6 terabytes of data in a massive data-driven approach to characterize this problem fully. We have developed a full characterization of the three-dimensional space that characterizes these games expect in a region near the plane x+y=z. In the past several months, we have uncovered more about this region and have started work to characterize individual games in the space more precisely than previously thought possible. This material is supported by the NSF grant 1246818 and by the NSF-supported REUF program of the American Institute of Mathematics.

Combinatorial Formulas for Restricted Stirling and Lah Number Matrices and their Inverses

Clifford Smyth
UNC Greensboro
cdsmyth@uncg.edu
Coauthors: John Engbers (Marquette University) David Galvin (University of Notre Dame)

Given a set R of natural numbers let S(n,k,R) be the restricted Stirling number of the second kind: the number of ways of partitioning a set of size n into k non-empty subsets with the sizes of these subsets restricted to lie in R. Let S(R) be the matrix with S(n,k,R) in its (n,k) entry. If R contains 1, S(R) has an inverse T(R) with integer entries. We find that for many R the entries T(n,k,R) of T(R) are expressible (up to sign) as the cardinalities of explicitly defined sets of trees and forests. For example this is the case when R has no exposed odds, i.e. R contains 1 and 2 and R never contains an odd number n greater than 1 without also containing n+1 and n-1. We have similar results for restricted Stirling numbers of the first kind (partitions into cycles) and Lah numbers (partitions into ordered lists). Our proofs depend in part on a combinatorial formula for the coefficients of the compositional inverse of a power series that expresses each coefficient as a sum of weighted trees.

Deep Learning for Spatio-Temporal Modeling

Vadim Sokolov
George Mason University
vsokolov@gmu.edu
Coauthors: Matthew F Dixon, Nicholas G Polson

Deep learning applies hierarchical layers of hidden variables to construct nonlinear high dimensional predictors. Our goal is to develop and train deep learning architectures for spatio-temporal modeling. Training a deep architecture is achieved by stochastic gradient descent (SGD) and drop-out (DO) for parameter regularization with a goal of minimizing out-of-sample predictive mean squared error. To illustrate our methodology, we predict the sharp discontinuities in traffic flow data, and secondly, we develop a classification rule to predict short-term futures market prices as a function of the order book depth. Finally, we conclude with directions for future research.

Stationary Yield to Maturity Zero Coupon Bonds Historical Simulation Value at Risk

J. Beleza Sousa
CMA-ISEL Portugal
jsousa@deetc.isel.ipl.pt
Coauthors: Manuel L. Esquível, Raquel M. Gaspar

Due to bond prices pull to par, zero coupon bonds historical returns are not stationary as they tend to zero as time to maturity approaches. Given that the historical simulation method for computing Value at Risk (VaR) requires a stationary sequence of historical returns, zero coupon bonds historical returns can not be used to compute Value at Risk (VaR) by historical simulation. Their use would systematically overestimate VaR, resulting in an invalid VaR sequence. In this paper we propose an adjustment of zero coupon bonds historical returns, that allows computing VaR by historical simulation. We prove that the proposed adjustment applies whenever the zero coupon bonds continuously compounded yields to maturity are stationary. We illustrate the VaR computation with historical returns and with adjusted historical returns, in a simulation scenario.

Sequential Design and Analysis of Mixture Experiments based on Gaussian Processes

Jon Stallings
North Carolina State University
jwstalli@ncsu.edu
Coauthors: Munir Winkel

Conventional data analysis of a mixture experiment is based on polynomial models constrained to account for the fact that the components sum to 1. Similar to non-mixture response surface situations, this approach is reasonable if one considers a restricted design space of the entire design simplex. A cubic polynomial analysis of a full-simplex design is often recommended, but may still be too simplistic. If a polynomial model is implemented in a restricted space, one may find that the space does not contain the optimum mixture settings, and the constrained model complicates standard sequential design approaches. We propose here the use of Gaussian process (GP) models to analyze mixture data from physical experiments and demonstrate how the model leads to a clear sequential design approach. In addition to avoiding the technical issues found with mixture polynomial models, we show GP models possess superior prediction properties for mixture experiments than predictions based on a polynomial model.

Information-Based Subdata Selection

John Stufken
Arizona State University
jstufken@asu.edu
Coauthors: HaiYing Wang; Min Yang

Simply due to size, in order to analyze a huge data set, it may be necessary or desirable to perform the analysis on selected subdata. There are various methods for selecting subdata from big data, including sampling-based methods and methods that advocate the use of information-based criteria. The information-based criteria relate the problem of "optimal" subdata selection to the problem of optimal design of experiments. While there are significant differences between the two problems, the connection makes tools from optimal design available for subdata selection problems. We introduce the basic ideas, demonstrate the success of information-based methods, and discuss some of the shortcomings.

Multivariate Association Test for Rare Variant Controlling for Cryptic and Family Relatedness

Jianping Sun
University of North Carolina at Greensboro
j_sun4@uncg.edu
Coauthors: Karim Oualkacha, Celia Greenwood, and Lajmi Lakhal-Chaieb.

In genetic studies of complex diseases, multiple measures of related phenotypes are often collected. Jointly analyzing these phenotypes may improve power to detect sets of rare variants affecting multiple traits. In this work, we consider association testing between a set of rare variants and multiple phenotypes in family-based designs. We use a mixed linear model to express the correlations among the phenotypes and between related individuals. Given the many sources of correlations in this situation, deriving an appropriate test statistic is not straightforward. We derive a vector of score statistics, whose joint distribution is approximated using a copula. This allows us to have closed-form expressions for the p-values of several test statistics. A comprehensive simulation study and an application to Genetic Analysis Workshop 18 (GAW18) data highlight the gains associated with joint testing over univariate approaches, especially in the presence of pleiotropy or highly correlated phenotypes.

Bayesian Factor Analysis Regression with Incorporation of Grouping Information

Thierry Chekouo T.
University of Calgary
thierry.chekouotekou@ucalgary.ca
Coauthors: Sandra Safo

Recent advances in data collection and processing in biomedical research allow different data types to be measured on the same subjects, with each data type measuring different sets of characteristics, but collectively helping to explain underlying complex mechanisms. In some instances, phenotypic data are also available. The main goals of these problems are to study the overall dependency structure among the data types, and to develop a model for predicting future phenotypes. Canonical correlation analysis is oftentimes used for such problems. We present a Bayesian canonical correlation framework that simultaneously models the overall association between data types using only relevant variables, while also predicting future outcomes using the canonical correlation variates. In addition, through prior distributions, we incorporate in our model prior structural information (such as biological networks) within each data type that allows us to select functionally meaningful networks involved in the determination of canonical correlation variates. We demonstrate the effectiveness of the proposed approach using simulations and observed data.

Variable Selection in Mixture Models: Uncovering Cluster Structure and Relevant Features

Mahlet Tadesse
Department of Mathematics and Statistics, Georgetown University
mgt26@georgetown.edu

Uncovering cluster structure and identifying relevant features can shed important insights when analyzing high-dimensional data. In this talk, I will present methods we have proposed to address this problem in a unified manner. I will start by discussing variable selection in the context of unsupervised clustering, where the goal is to uncover the latent classes while identifying variables that discriminate between the different groups. This may consist, for example, in using genomic data to simultaneously discover disease subtypes and locate markers that distinguish between these subtypes. In the second part of the talk, I will focus on the problem of relating two high-dimensional data sets, as in integrative genomic studies, where there is interest in finding relationships between genomic data from different sources. I will discuss methods we have proposed that combine ideas of mixture of regression models and variable selection to uncover correlated response profiles and identify cluster-specific subsets of covariates. I will illustrate the methods with various applications.

Classifying Hate Speech Using a Two-Stage Model

Yiwen Tang
Wake Forest University
tangy215@wfu.edu
Coauthors: Nicole Dalzell (Wake Forest University)

Social media and other online sites are being increasingly scrutinized as platforms for cyber bullying and hate speech. Many machine learning algorithms, such as support vector machines, have been used to create classification tools to identify and potentially filter such patterns of negative speech. While effective for prediction, these methodologies yield models that are difficult to interpret. They also tend to focus on classifying comments as either negative or not, rather than separating negative comments into categories like hate speech or speech indicating a threat. To address both of these concerns, we introduce a two-stage classification model. The first stage incorporates a machine learning algorithm to classify each comment as negative or neutral. The second stage generates a frequency-based internal lexicon from a pre-classified training data set. The training data are classified as some combination of toxic, severe toxic, obscene, threat, insult, and identity hate. The lexicon is used to assign each comment in the test data set a score corresponding to each category. These scores are then used to classify the comments. We illustrate our approach using a large data set of comments from Wikipedia.

Non-consistency of MOM Variogram Estimators on the Sphere

Romesh Ruwan Thanuja
University of North Carolina at Greensboro
r_athuru@uncg.edu

The Variogram estimator is commonly used for characterizing spatial dependency in spatial statistics. It has been known that the Method of Moments (MOM) variogram estimator is not consistent when the spatial process on the circle is stationary. In this research, we extend this result on the sphere when the underlying axially symmetric process is longitudinally reversible. Some initial results have been obtained and a simulation study will be performed to validate our findings.

Computing the Effect of Measurement Errors on Efficient Variant of the Product and Ratio Estimators of Mean Using Auxiliary Information

Gajendra K. Vishwakarma
Department of Applied Mathematics, Indian Institute of Technology (ISM) Dhanbad, Dhanbad-826004, India
vishwagk@rediffmail.com
Coauthors: Neha Singh and Jong Min Kim

This paper presents an efficient variant of the usual product and ratio methods of estimation of population mean of a study variable Y in the context of simple random sampling when the observations of both study variables and auxiliary variables are commingled with measurement error. The mean square error of proposed estimator has been derived and studied under measurement error. Monte-Carlo simulation and numerical studies have been carried out to study the properties of the estimators and compared with the mean square error and percentage relative efficiency of the estimator when variables are free from measurement errors. Keywords: Design parameter, Measurement error, Percentage Relative Efficiency, Monte- Carlo Simulation.

Using Statistics to Solve Biological Problems: An Example of Termite Recombination

Prashant Waiker
University of North Carolina at Greensboro
p_waiker@uncg.edu
Coauthors: Edward Vargo, Texas A&M University, TX, USA; Kenji Matsuura, Kyoto University, Kyoto, Japan; Paul Labadie, NC State University, NC, USA; Olav Rueppell, University of North Carolina at Greensboro, NC, USA

Meiotic recombination is a universal biological process in all sexual organisms. It results in a reciprocal exchange of genetic information between the genomic strands which increases genetic diversity to produce new genomic combinations. The recombination rate is highly variable across species and causes and consequences of which is not fully understood yet. In social insects (bees, wasps, ants, and termites), one of the popular hypothesis argues that high recombination rates are beneficial for eusociality because they increase genetic diversity and disease resistance. Empirical evidence from bees, ants, and wasps have supported this hypothesis, but no study is available on termites. Moreover, the recent theoretical evidence contradicts the hypothesis, and more data is needed. A useful source of information regarding recombination events can be found in population genetic data, and recent advent in genotyping and sequencing technologies have greatly increased the availability of such data. However, interpretation of sequence patterns can be problematic without an understanding of the process that generated the patterns. Statistical modeling of the evolutionary process by which the data was generated can, therefore, provide a useful tool by which patterns of variation can be understood. In this talk, I am going to introduce different aspects related to the estimation of meiotic recombination using application of statistics and present an example of analysis on a group of social insects-termites. In this study, we genotyped 96 and 77 children from parents of two different species of termites and constructed linkage maps using quantitative analysis R package- RQTL to determine recombination rates of termites. We found unlike other social insects, termites do not show exceptional recombination rates, and thus we reject the hypothesis that states eusociality may be a consequence of high recombination.

Optimal Maximin L1-Distance Latin hypercube Designs Based on Good Lattice Point Designs

Lin Wang
UCLA
linhappyforever@ucla.edu
Coauthors: Qian Xiao, Hongquan Xu

Maximin distance Latin hypercube designs are commonly used for computer experiments, but the construction of such designs is challenging. We construct a series of maximin Latin hypercube designs via Williams transformations of good lattice point designs. Some constructed designs are optimal under the maximin L1-distance criterion, while others are asymptotically optimal. Moreover, these designs are also shown to have small pairwise correlations between columns.

Integrative Survival Analysis with Uncertain Event Times in Application to a Suicide Risk Study

Wenjie Wang
University of Connecticut
wenjie.2.wang@uconn.edu
Coauthors: Robert Aseltine, Kun Chen, and Jun Yan

The concept of integrating data from disparate sources to accelerate scientific discovery has generated tremendous excitement in many fields. The potential benefits from data integration, however, may be compromised by the uncertainty due to incomplete/imperfect record linkage. Motivated by a suicide risk study, we propose an approach for analyzing survival data with uncertain event times arising from data integration. Specifically, in our problem deaths identified from the hospital discharge records together with

reported suicidal deaths determined by the Office of Medical Examiner may still not include all the death events of patients, and the missing deaths can be recovered from a complete database of death records. Since the hospital discharge data can only be linked to the death record data by matching basic patient characteristics, a patient with a censored death time from the first dataset could be linked to multiple potential event records in the second dataset. We develop an integrative Cox proportional hazards regression, in which the uncertainty in the matched event times is modeled probabilistically. The estimation procedure combines the ideas of profile likelihood and the expectation conditional maximization algorithm (ECM). Simulation studies demonstrate that under realistic settings of imperfect data linkage, the proposed method outperforms several competing approaches including multiple imputation. A marginal screening analysis using the proposed integrative Cox model is performed to identify risk factors associated with death following suicide-related hospitalization in Connecticut. The identified diagnostics codes are consistent with existing literature and provide several new insights on suicide risk prediction and prevention.

Minimum Entropy Clustering of Functional Data

Erik L. Wendt
Gettysburg College
wender01@gettysburg.edu
Coauthors: Kumer P. Das

Minimum entropy clustering, which uses the entropy of clusters as a clustering criterion, is a recent developed clustering method. While there is a lot of work on applying entropy as a clustering criterion of discrete data, there is no work on using entropy as a clustering criterion for functional data. In this talk, we present criteria for minimum entropy clustering of functional data and compare it with other functional clustering methods, such as functional k-means clustering.

Modeling Daily Seasonality of Mexico City Ozone using Nonseparable Covariance Models on Circles Cross Time

Philip White
Duke University
paw27@duke.edu
Coauthors: Emilio Porcu

Mexico City tracks ground-level ozone levels to assess compliance with national ambient air quality standards and to prevent environmental health emergencies. Ozone levels show distinct daily patterns, within the city, and over the course of the year. To model these data, we use covariance models over space, circular time, and linear time. We review existing models and develop new classes of nonseparable covariance models of this type, models appropriate for quasi-periodic data collected at many locations. With these covariance models, we use nearest-neighbor Gaussian processes to predict hourly ozone levels at unobserved locations in April and May, the peak ozone season, to infer compliance to Mexican air quality standards and to estimate respiratory health risk associated with ozone. Predicted compliance with air quality standards and estimated respiratory health risk vary greatly over space and time. In some regions, we predict exceedance of national standards for more than a third of the hours in April and May. On many days, we predict that nearly all of Mexico City exceeds nationally legislated ozone thresholds at least once. In peak regions, we estimate respiratory risk for ozone to be 55% higher on average than the annual average risk and as much at 170% higher on some days.

Non-penalized Variable Selection via Generalized Fiducial Inference

Jonathan P Williams
University of North Carolina at Chapel Hill
jpwill@live.unc.edu
Coauthors: Jan Hannig

Standard penalized methods of variable selection and parameter estimation rely on the magnitude of coefficient estimates to decide which variables to include in the final model. However, coefficient estimates are unreliable when the design matrix is collinear. To overcome this challenge an entirely new perspective on variable selection is presented within a generalized fiducial inference framework. This new procedure is able to effectively account for linear dependencies among subsets of covariates in a high-dimensional setting where p can grow almost exponentially in n, as well as in the classical setting where $p \leq n$. It is shown that the procedure very naturally assigns small probabilities to subsets of covariates which include redundancies by way of explicit L_0 minimization. Furthermore, with a typical sparsity assumption, it is shown that the proposed method is consistent in the sense that the probability of the true sparse subset of covariates converges in probability to 1 as $n \to \infty$, or as $n \to \infty$ and $p \to \infty$. Very reasonable conditions are needed, and little restriction is placed on the class of possible subsets of covariates to achieve this consistency result.

Fractional Factorial Designs with Clear Two-Factor Interactions

Huaiqing Wu
Iowa State University
isuhwu@iastate.edu
Coauthors: Robert Mee; Boxin Tang

Regular two-level fractional factorial designs are commonly used to identify important factors. We consider the problem of selecting such designs that allow joint estimation of all main effects and some specified two-factor interactions (2FIs) without aliasing from other 2FIs. We address the general problem by finding, among all $2^{(m-p)}$ designs with given m and p, those resolution IV designs whose sets of clear 2FIs contain the specified 2FIs as subsets. We use a linear graph to represent the set of clear 2FIs for a resolution IV design, where each line connecting two vertexes represents a clear 2FI between the factors represented by the two vertexes. We call a $2^{(m-p)}$ resolution IV design admissible if its graph is not isomorphic to any proper subgraph of the graph of any other $2^{(m-p)}$ resolution IV design. We show that all even resolution IV designs are inadmissible. We then use a classical subgraph-isomorphism algorithm to determine all admissible designs of 32, 64, and 128 runs. This leads to a concise catalog of all admissible designs of 32 and 64 runs, and a lengthy but substantially reduced list (compared with the number of non-isomorphic designs) for 128 runs.

Machine Learning for Music Mining with LDA Model

Qiuyi Wu
Rochester Institute of Technology
qw9477@rit.edu
Coauthors: Ernest Fokoue

Extensive studies have been conducted on both musical scores and audio tracks of western classical music with the finality of learning and detecting the key in which a particular piece of music was played. Both the Bayesian Approach and modern unsupervised learning via latent Dirichlet allocation have been used for such learning tasks. In this research work, we venture out of the western classical genre and embrace and explore jazz music. We consider the musical score sheets and audio tracks of some of the giants of jazz like Duke Ellington, Miles Davis, John Coltrane, Dizzie Gillespie, Wes Montgomery, Charlie Parker, Sonny Rollins, Louis Armstrong (Instrumental), Bill Evans, Dave Brubeck, Thelonious Monk (Pianist). We specifically employ Bayesian techniques and modern topic modelling methods (and even occasionally a combination of

both) to explore tasks such as: automatic improvisation detection, genre identification, key learning (how many keys do the giants of jazz tended to play in, and what are those keys) and even elements of the mood of the piece.

Borel-Tanner Distribution and Bayes Estimators for the Basic Reproduction Number of an Epidemic

George Yanev
The University of Texas Rio Grande Valley
`george.yanev@utrgv.edu`

We construct a monotone version of an empirical Bayes estimator for the parameter of Borel-Tanner (BT) distribution. The BT distribution arises, for example, in branching stochastic processes and queueing theory. Our interest stems from its role as the distribution of the total number of infected individuals in an epidemic modeled by a branching process.

Information-Based Optimal Subdata Selection for Mixture Modelling

Min Yang
University of Illinois at Chicago
`minyang.stat@gmail.com`

How to implement data reduction to draw useful information from big data is a hot spot of modern scientific research. One attractive approach is data reduction through subdata selection. Typically, this approach is based on some strong model assumption: data follows one specific model. Big data is complexity and it may not be the best to model the data using a specific model. A better approach to describe the data is through mixture modelling. How to select informative subdata under mixture modelling? In this talk, a new framework is proposed to address this issue.

Newer Variations of the Unrelated Question Binary RRT Model Examining the Impact of Untruthful Responding

Amber Young
Department of Statistics, Purdue University, West Lafayette, IN
`young268@purdue.edu`
Coauthors: Ryan Parks and Sat Gupta, Department of Mathematics and Statistics, University of North Carolina at Greensboro

Estimating the prevalence of a sensitive trait in a population is not a simple task due to the general tendency among survey respondents to answer sensitive questions in a way that is socially desirable. Use of Randomized Response Techniques (RRT) is one of several approaches for reducing the impact of this tendency. We propose using an inverse sampling based optional unrelated-question RRT model. We observe that while optionality does lead to more efficient estimation, the use of inverse sampling does not necessarily do so. However, inverse sampling is still useful when prevalence of the sensitive trait is low by providing an extra layer of precaution. We also consider the impact of untruthful responding on our model and observe that even if only a small number of respondents lie, the model efficiency decreases. This emphasizes the importance of pre-survey respondent training. Our results are validated using both theoretical comparisons and computer simulations.

A Data-driven Approach to Predicting Diabetes and Cardiovascular Disease with Machine Learning

Stacey Miertschin (Winona State University) and Amber Young (Purdue University)
UNCG Statistics REU
smiertschin14@winona.edu
Coauthors: An Dinh, Somya Mohanty

Diabetes and cardiovascular disease are two of the main causes of death in the United States. Our research explores a data-driven approach of using supervised machine learning models to predict these diseases using the NHANES dataset. We evaluate performance characteristics of different models applied towards predicting diabetes, prediabetes, and cardiovascular disease in a patient. The paper also explores a weighted ensemble model which combines the results of different models for higher predictability. Using an iterative approach based on an ensemble of trees, we identify the key features in predicting diabetes and cardiovascular disease. Within the approach we compare and contrast models across multiple datasets based on the availability of features and observations.We conclude that a data-driven approach can lead to higher accuracy in predictability.

Intrinsic Random Functions and Universal Kriging On the Circle

Haimeng Zhang
UNC Greensboro
h_zhang5@uncg.edu

A common assumption made when describing spatial dependency is second order stationarity, that is, the mean of the spatial process on the circle is constant and the covariance function at any two points depends only on their angular distance. However, this assumption is often not satisfied in practice. In order to model non-stationary processes on the circle, we extend the notion of intrinsic random functions and show that low-frequency truncation plays an essential role. We present these developments using the theory of reproducing kernel Hilbert space and further discuss the link between universal kriging and splines.

Comparison of Mean Estimators of Sensitive Variables under Measurement Errors with Respect to Efficiency and Respondent Privacy

Qi Zhang
Department of Mathematics and Statistics, UNC Greensboro
q_zhang@uncg.edu
Coauthors: Sat Gupta (Department of Mathematics and Statistics, UNC Greensboro), Sadia Khalil (Department of Statistics, Lahore College for Women University, Lahore, Pakistan)

In this study, our primary focus is on examining if using optional RRT models as opposed to non-optional RRT models for mean estimation of a sensitive variable when measurement errors are present, produces more efficient estimators. A unified measure of model efficiency and respondent privacy will be used in this comparison.

Design Based Incomplete U-statistics

Wei Zheng
university of tennessee
wzheng9@utk.edu
Coauthors: Xiangshun Kong

Many statistical and machine learning problems are formulated as a U-statistic. The statistical properties of variants of U-statistics has been extensively. Meanwhile, its computational cost is prohibitively expensive especially with the increasing size of data in modern applications. Even a moderate size data could be large enough to present a huge challenge for computing. In this talk, I will share some insight on how the design methods could help alleviate the computational issue while maintaining the minimum variance property of U-statistics.

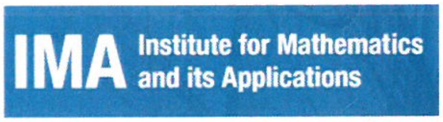
Founding Program Director
Master of Science in Informatics & Analytics Program

The University of North Carolina at Greensboro seeks a dynamic Founding Director on a 12-month tenure-track or tenured appointment for its newly established Master of Science in Informatics & Analytics program. The preferred starting date is August 1, 2019. The MSI&A is a transdisciplinary applied program intersecting multiple academic domains related to analytics and data science. The position is open only to candidates holding a Ph.D. degree, with work experience in the fields of analytics and data science, and, who have at least a five-year post-Ph.D. experience. Candidates from academia must currently hold a rank of at least an Associate Professor. Candidates from Industry may be considered if they have at least five-years of academic experience.

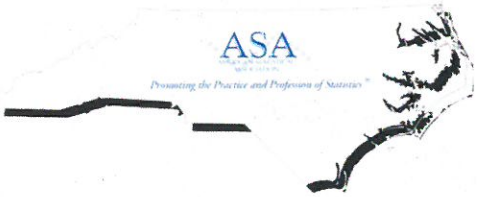
A strong record of teaching and research at the university level with an understanding of and commitment to teaching within a culturally diverse environment is essential. The successful candidate will be primarily expected to provide leadership and general administration of the MSI&A program, teach courses in the graduate core and/or the candidate's disciplinary concentration, as well as maintain an active research program.

Application materials should be submitted electronically at `http://jobsearch.uncg.edu`. Applications should include a cover letter, curriculum vita, description of current and past research, a statement of teaching philosophy, and a statement describing their Informatics & Analytics background. Applicants are also asked to provide the names, email addresses, and phone numbers of three (3) professional References (at least one of these references should address the candidate's background in Informatics & Analytics). Review of applications will begin on November 1, 2018 and will continue until the position is filled. For specific questions regarding this position, please contact the Interim Program Director & Search Chair Dr. Sat Gupta at msia@uncg.edu

UNCG is a Minority Serving doctorate-granting Institution that prides itself in both faculty and student diversity and we seek to attract a diverse applicant pool for this position. UNCG is an EOE//M/F/D/V employer and are strongly committed to increasing faculty diversity.


IMA
Institute for Mathematics and its Applications

1 8 9 1
UNC
GREENSBORO


ASA
Promoting the Practice and Profession of Statistics

ASA
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics

Pearson

sas

NISS
National Institute of Statistical Sciences

Springer

Rho
Giving flight to research