AISC–2022 International Conference on Advances in Interdisciplinary Statistics and Combinatorics

October 7-9, 2022

The University of North Carolina at Greensboro

Contents

Conference Sponsors

The University of North Carolina at Greensboro

Institute for Mathematics and Its Applications

National Institute of Statistical Sciences

Springer

International Conference on
Advances in Interdisciplinary Statistics and Combinatorics
AISC – 2022

Local Organizing Committee

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

John Stufken (Co-Chair)
George Mason University

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

Scott Richter
Department of Mathematics and Statistics, UNCG

Xiaoli Gao
Department of Mathematics and Statistics, UNCG

Igor Erovenko
Department of Mathematics and Statistics, UNCG

Jianping Sun
Department of Mathematics and Statistics, UNCG

David Bickel
Informatics and Analytics Program, UNCG

Scientific Program Committee

Sat Gupta (Chair)
Department of Mathematics and Statistics
UNCG

John Stufken (Co-Chair)
George Mason University

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

Xiaoli Gao
Department of Mathematics and Statistics
UNCG

Scott Richter
Department of Mathematics and Statistics
UNCG

Jianping Sun
Department of Mathematics and Statistics
UNCG

David Banks
Department of Statistical Science
Duke University

Hrishikesh Chakraborty
Department of Biostatistics and Bioinformatics
Duke University

Sujit Ghosh
Department of Statistics
NC State University

Angela Dean
Department of Statistics
The Ohio State University

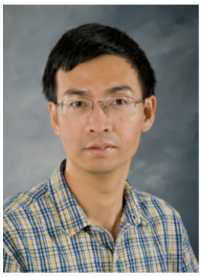
Benjamin Kedem
Department of Mathematics
University of Maryland

Abhyuday Mandal
University of Georgia

Arman Sabbaghi
Purdue University

Sadia Khalil
Lahore College for Women University, Lahore, Pakistan

Welcome from the Organizers


On behalf of The University of North Carolina at Greensboro, we are excited to welcome you to the International Conference on *Advances in Interdisciplinary Statistics and Combinatorics (AISC-2022)*. Normally, AISC conferences are held during even years, but we could not do it in 2020 due to Covid. We managed to do one in 2021 but it was only a virtual one. We are happy that we are finally able to greet you in person.

We would like to thank all the external sponsors of this conference including the Institute for Mathematics and its Applications (IMA), National Institute of Statistical Sciences (NISS), and Springer. Additionally, we would like to recognize local support from UNC Greensboro including Dr. John Z. Kiss, Dean of the College of Arts and Sciences; the conference secretaries Leslie Justice and Dafne Sanchez; and our System Administrator Richard Cheek. We also want to thank the very dedicated group of volunteers (all of them PhD students in the Department) who have helped tremendously, particularly Wendy Shou who maintained the conference website and prepared the source file for the conference book of abstracts. We also want to thank all of the plenary speakers, the session organizers, and the presenters whose contributions have enriched the conference program.

We invite you to browse this conference program book as well as the conference website (`https://sites.google.com/uncg.edu/aisc-2022/`) to know more about this conference as well as earlier AISC conferences. You will notice that this conference features 2 keynote addresses by Wendy Martinez (ASA President in 2020 and a recipient of the ASA Founders Award in 2017) and by Srinivasa Varadhan (Frank Jay Gould Professor of Science at Courant Institute/NYU, winner of the Abel Award (2007), and the winner of the National Medal of Science from President Obama (2010)). The conference also features 6 plenary talks and 125 other presentations spread over 29 parallel sessions. We would also like to draw your attention to a panel discussion on Statistical Consulting on Friday, October 7, from 6:00-7:20 pm in the EUC auditorium immediately preceding the conference reception. This panel discussion is organized and chaired by Dan Hall from the University of Georgia.

We thank you for choosing to attend the AISC-2022 conference and sincerely hope that your participation in the conference will be highly productive. Hopefully, your visit to our campus and the city of Greensboro will be a pleasant experience.

Sat Gupta
Conference Chair

Haimeng Zhang & John Stufken
Conference Co-Chairs

Plenary Speakers: AISC 2022

Keynote Address

Wendy Martinez
ASA President 2020
Senior Mathematical Statistician for Data Science
U.S. Census Bureau

Wendy Martinez is the Senior Mathematical Statistician for Data Science in the Research and Methodology Directorate at the US Census Bureau. Prior to this time, she served as the Director of the Mathematical Statistics Research Center at the Bureau of Labor Statistics (BLS) and worked in several research positions throughout the Department of Defense, including the position of Science and Technology Program Officer at the Office of Naval Research. She is the lead author of three books on MATLAB and statistics, the Coordinating Editor of Statistics Surveys, and the Editor of Chance magazine starting in 2023.
Wendy received her Ph.D. in Computational Sciences and Informatics with an emphasis on computational statistics at George Mason University in 1995. She completed a Master of Science degree in engineering at George Washington University in 1991 and a Bachelor of Science degree in mathematics and physics at Cameron University in 1989. She also received a Graduate Certificate in Survey Statistics from the University of Maryland in 2015. Wendy was elected as a Fellow of the American Statistical Association (ASA) in 2006 and is an elected member of the International Statistical Institute. She received the ASA Founders Award in 2017 and is honored to have served as the 2020 ASA President.
Wendy was elected as the 2023 chair of the ASA Justice Equity Diversity and Inclusion (JEDI) Outreach Group and the ASA Text Analysis Interest Group. She is a member of the Federal Committee on Statistical Methodology, where she established and leads two interest groups – one on geospatial data and one on computational statistics. She is the co-founder of R Govys (supported by the R Consortium) and leads the Inter-agency R Users Group.

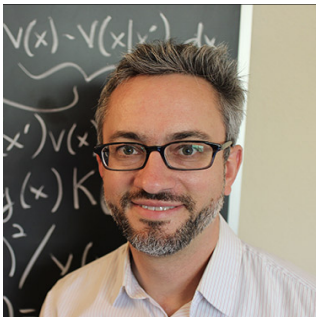
Srinivasa Varadhan
NYU/Courant Institute

Varadhan is currently Frank Jay Gould Professor of Science at the Courant Institute. He is known for his work with Daniel W Stroock on diffusion processes, and for his work on large deviations with Monroe D Donsker.Varadhan's awards and honors include the National Medal of Science from President Barack Obama. He also received the Birkhoff Prize and the Leroy P Steele Prize for Seminal Contribution to Research from the American Mathematical Society, awarded for his work with Daniel W Stroock on diffusion processes. He was awarded the Abel Prize in 2007 for his work on large deviations with Monroe D Donsker.The Government of India awarded him the Padma Bhushan. He has received honorary degrees from Université Pierre et Marie Curie in Paris, from Indian Statistical Institute, Chennai Mathematical Institute and Duke University.
Varadhan is a member of the US National Academy of Sciences and the Norwegian Academy of Science and Letters. He was elected Fellow of the American Academy of Arts and Sciences, the Third World Academy of Sciences, the Institute of Mathematical Statistics, the Royal Society, the Indian Academy of Sciences, the Society for Industrial and Applied Mathematics and the American Mathematical Society. More information about Professor Varadhan is available at `https://math.nyu.edu/~varadhan/`.

Plenary Speakers

Richard Davis
Columbia University

Richard Davis is the Howard Levene Professor of Statistics at Columbia University and former chair of the Statistics Department (2013-19). He has held academic positions at MIT, Colorado State University, and visiting appointments at numerous other universities. He was Hans Fischer Senior Fellow at the Technical University of Munich (2009-12) , Villum Kan Rasmussen Visiting Professor (2011-13) at the University of Copenhagen , and Jubilee Professor at Chalmers University (2019). Davis is a fellow of the Institute of Mathematical Statistics and the American Statistical Association, and is an elected member of the International Statistical Institute. He was president of IMS in 2016 and Editor-in-Chief of *Bernoulli Journal* 2010-12. He is co-author (with Peter Brockwell) of the best-selling books, Time Series: Theory and Methods, Introduction to Time Series and Forecasting, and the time series analysis computer software package, ITSM2000. Together with Torben Andersen, Jens-Peter Kreiss, and Thomas Mikosch, he co-edited the Handbook in Financial Time Series and with Holan, Lund, and Ravishanker, the book, Handbook of Discrete-Valued Time Series. In 1998, he won (with collaborator W.T.M Dunsmuir) the Koopmans Prize for Econometric Theory. He has advised/co-advised 34 PhD students. His research interests include time series, applied probability, extreme value theory, and spatial-temporal modeling. More information about Professor Davis is available at `http://www.stat.columbia.edu/~rdavis/`.



Bobby Gramacy
Virginia Tech

Bobby Gramacy is a Professor of Statistics in the College of Science at Virginia Polytechnic and State University (Virginia Tech/VT) and affiliate faculty in VT's Computational Modeling and Data Analytics program. Previously he was an Associate Professor of Econometrics and Statistics at the Booth School of Business, and a fellow of the Computation Institute at The University of Chicago. Bobby's research interests include Bayesian modeling methodology, statistical computing, Monte Carlo inference, nonparametric regression, sequential design, and optimization under uncertainty.

Bobby Gramacy is a computational statistician. Bobby specializes in areas of real-data analysis where the ideal modeling apparatus is impractical, or where the current solutions are inefficient and thus skimp on fidelity. Such endeavors often require new models, new methods, and new algorithms. His goal is to be impactful in all three areas while remaining grounded in the needs of a motivating application. Bobby aims to release general purpose software for consumption by the scientific community at large, not only other statisticians. More information about Professor Gramacy is available at `https://bobby.gramacy.com/`.

Eiren Jacobson
Centre for Research into Ecological and Environmental Modelling, University of St Andrews, UK

Eiren Jacobson is a Research Fellow at the University of St Andrews, where she works on quantitative approaches to population studies of marine mammals.
Eiren grew up in Oregon and received her BA in Environmental Biology from Columbia College. She then worked as a technician at the National Oceanic and Atmospheric Administration Southwest Fisheries Science Center before undertaking a PhD in Biological Oceanography at the Scripps Institution of Oceanography. For her doctoral dissertation, Eiren worked in collaboration with scientists at the NOAA SWFSC to develop a passive acoustic monitoring network for harbor porpoise in Monterey Bay, California. Eiren was a Postdoctoral Research Associate at the University of Washington School of Aquatic and Fishery Sciences, where she collaborated with scientists at the NOAA Alaska Fisheries Science Center on an integrated population model for Cook Inlet beluga whales.
Since 2018, Eiren has lived in Scotland and worked at the Centre for Research into Ecological and Environmental Modelling at the University of St Andrews. There, she has contributed to multiple research projects, including quantifying the impacts of Naval sonar on Blainville's beaked whales in Hawaii, developing an integrated population model to investigate the decline of harbor seal populations in Scotland, improving methods to estimate grey seal pup production in the UK, and assessing the statistical power of visual and passive acoustic surveys to detect trends in marine mammal populations. In her spare time, Eiren enjoys sewing, reading, board games, and spending time with her dogs Juniper and Phoenix, cat Fern, and partner Dave. More information about Professor Jacobson is available at `http://eirenjacobson.info/`.

Karen Kafadar
ASA President 2019
University of Virginia

Karen Kafadar is Commonwealth Professor & Past Chair of Statistics at the University of Virginia. She received her BS & MS from Stanford and her PhD from Princeton, and is a Fellow of the ASA AAAS, and ISI for her research on statistical methods & data analysis in the physical, chemical, biological, and engineering sciences, and for service on scientific panels. She is past Editor of JASA Reviews, Technometrics and most recently Editor-in-Chief of The Annals of Applied Statistics, and was 2019 ASA President. Her most recent work concerns statistical methodology for problems in eyewitness identification and for randomized cancer screening trials, the subject of her talk here today. More information about Professor Kafadar is available at `https://statistics.as.virginia.edu/faculty-staff/profile/kk3ab`.

Regina Liu
Rutgers University

Regina Liu is currently Distinguished Professor of Statistics at Rutgers University. She received her PhD in statistics from Columbia University. Her research areas include data depth, resampling, confidence distribution, and fusion learning. Aside from theoretical and methodological research, she has long collaborated with the FAA on aviation safety research projects on statistical process control, text mining and risk management. She has served as Editor for the Journal of the American Statistical Association and the Journal of Multivariate Analysis, and as Associate Editor for several journals, including the Annals of Statistics. She is an elected fellow of the Institute of Mathematical Statistics and the American Statistical Association, and an elected member of the International Statistical Institute. Among other distinctions, she is the recipient of 2021 Noether Distinguished Scholar Award (American Statistical Association), 2011 Stieltjes Professorship (Thomas Stieltjes Institute for Mathematics, The Netherlands), and has delivered an IMS Medallion Lecture among other named lectures. She was elected President of the Institute of Mathematical Statistics, 2020-2021. More information about Professor Liu is available at `https://statistics.rutgers.edu/people-pages/faculty/people/130-faculty/391-regina-y-liu`.

Jerry Reiter
Duke University

Jerry Reiter is Professor of Statistical Science and Dean of the Natural Sciences at Duke University. His primary areas of research include methods for handling missing and erroneous values, for ensuring data privacy, for combining information across sources, and for analyzing complex data in the social sciences and public policy. He is a Fellow of the American Statistical Association and a Fellow of the Institute of Mathematical Statistics. He is the recipient of several teaching and mentoring awards from Duke University, including the Alumni Distinguished Undergraduate Teaching Award, the Outstanding Postdoctoral Mentor Award, and the Master's of Interdisciplinary Data Science Distinguished Faculty Award. He has advised multiple government agencies on creating data products to share with the public, as well as served on multiple panels and committees for the National Academy of Sciences. He received a Ph.D. in statistics from Harvard University in 1999. More information about Professor Reiter is available at `https://www2.stat.duke.edu/~jerry/`.

Conference Program
AISC 2022: October 7–9, 2022; UNC Greensboro

October 6, 2022, Thursday

4:00 – 7:00 pm Registration Desk, Azalea Room at the conference hotel, Holiday Inn, Gate City Blvd. You can pick your conference material here or at EUC on the 7^{th} morning.

October 7, 2022, Friday

7:45 – 8:15 Shuttles from Holiday Inn to UNCG
Departures from Holiday Inn: 7:45 am, 8:15 am
Please do not wait for the last shuttle. You may remain stranded and may have to take an Uber.

7:45 – Registration Desk, Refreshments, EUC Auditorium Lobby

9:00 – 9:15 Inaugural Remarks - EUC Auditorium
Sat Gupta – Conference Chair
Chuck Bolton – Associate Dean College of Arts and Sciences, UNCG

9:15 – 10:40 October 7, 2022, EUC Auditorium
Keynote Address I & Plenary Talk 1
Chair: Richard Davis – Columbia University
rdavis@stat.columbia.edu

9:15 – 10:00 Keynote Address
Wendy Martinez – ASA President (2020), US Census Bureau
What's the Big Deal With Data Ethics, and Why Should I Care?

10:00 – 10:40 Plenary Talk 1 Regina Liu, Rutgers University
Fusion learning: combine inferences from diverse data sources

10:40 – 11:00 Coffee Break EUC Lobby

11:00 – 1:00 October 7 - Parallel Session 1A: Sharpe Room EUC
Advances in Network Optimization
Chair: Patrick LeBlanc – Duke University
patrick.leblanc@duke.edu

David Banks - Duke University
Route Choice Under Uncertainty: An Adversarial Risk Analysis

Elvan Ceyhan - Auburn University
Comparison of Various Algorithms in Optimal Obstacle Placement with Disambiguation (OPD) Problem

Paul Zivich – UNC Chapel Hill
Targeted Maximum Likelihood Estimation With Network-Dependent Data

Justin Weltz – Duke University
Hidden Population Estimation With Auxiliary Information

11:00 – 1:00 **October 7 - Parallel Session 1B: Dail Room EUC**
Applications of Bayesian Statistical Methods
Chair: Jerry Reiter – Duke University
jerry@stat.duke.edu

Zeki Kazan – Duke University
Statistical Disclosure Risk with Differential Privacy, with Application to the 2020 Decennial Census

Emily Tallman – Duke University
Bayesian Predictive Decision Synthesis: Betting on Better Models

Rick Presman – Duke University
Inference for Distance-to-Set Regularization via Constrained Bayesian Inference

Alex Dombowsky – Duke University
Bayesian Multi-Study Clustering of Sepsis Patients: Utilizing Prior Information and Testing Cluster Discovery

11:00 – 1:00 **October 7 - Parallel Session 1C: Claxton Room EUC**
Topics in Big Data and Machine Learning
Chair: Kumer Pial Das – University of Louisiana Lafayette
kumer.das@louisiana.edu

Shaymal Halder - University of Pennsylvania
The Interplay of Biomass Energy Consumption On Ecological Footprint: Using Parametric and Time-Varying Non-Parametric Approaches

Joy D'Andrea - University of South Florida
A Brief Parametric Analysis of Catastrophic or Disastrous Hurricanes That Have Hit the Florida Keys between 1900 and 2000

Md Shamim Sarker – Radford University
Bayesian Pooled Testing Regression with Measurement Error

Mithun Acharjee - University of Alabama Birmingham
Statistical Shape Analysis of 3 DFN Data: Analysis Via Deformetrica

Kumer Das - University of Louisiana Lafayette
Air quality and lung cancer: Analysis via Local Control

11:00 – 1:00 **October 7 - Parallel Session 1D: Alexander Room EUC**
Advances in Active Learning
Organizers: Arman Sabbaghi & Timothy Keaton (Purdue University)
Chair: Arman Sabbaghi – Purdue University
sabbaghi@purdue.edu

Timothy Keaton - Purdue University
Actively Learning About Active Learning

Bill Heavlin – Google
Active Learning Meets Experimental Design Theory

Annie Sauer - Virginia Tech
Active Learning for Deep Gaussian Process Surrogates

Yubai Yuan - Pennsylvania State University
Query-Augmented Active Metric Learning

11:00 – 1:00 **October 7 - Parallel Session 1E: Kirkland Room EUC**
Modern Statistical Methods for Biomedical Studies
Chair: Zhezhen Jin, Columbia University
zj7@cumc.columbia.edu

Cuiling Wang - Albert Einstein College of Medicine
Optimal cut-point for disease incidence with censored data

Rui Wang - Harvard Medical School
Assessing exposure-time treatment effect heterogeneity in stepped wedge cluster randomized trials

Qin Shao - University of Toledo
Confidence Band Approach for Comparison of COVID19 Case Counts

Marianthi Markatou - University of Buffalo
Smoothing Kernels for Categorical and Mixed-Scale Data

11:00 – 1:00 **October 7 - Parallel Session 1F: Dogwood Room EUC**
Health Sciences Research
Chair: Jianping Sun – UNC Greensboro
j_sun4@uncg.edu

Jianping Sun – UNC Greensboro
Repeated Sampling in EMA Studies: A Discussion Statistical Challenge and Potential Solution

Grace Rhodes – Duke University
Markov Chain Composite Likelihood and Its Application in Genetic Recombination Model

Sunil Mathur – Houston Methodist, Weill Cornell Medical College
Cancer Data Science: Drug Testing in Cancer Research Using Auxiliary Information

Ciaran Evans – Wake Forest University
Two-Sample Testing With Local Community Depth

John Foley – Metron, Inc
Learning Communities in Data From Probabilistic Estimates of Similarity

1:00 – 2:00 **Lunch at Cone Ballrooms, EUC**

2:00 – 3:25 **October 7 - Plenary Talks 2/3: EUC Auditorium**
Chair: David Banks – Duke University
david.banks@duke.edu

2:00 – 2:40 **Jerry Reiter – Duke University**
How Auxiliary Information Can Help Your Missing Data Problem

2:45 – 3:25 **Robert Gramacy – Virginia Tech**
Deep Gaussian Process Surrogates for Computer Experiments

3:30 – 3:50 **Coffee Break**

3:50 – 5:50 **October 7 - Parallel Session 2A: Sharpe EUC**
Design of Experiments 1 – Recent Developments
Chair: Rakhi Singh – Binghamton University
rsingh@binghamton.edu

Arman Sabbaghi - Purdue University
A Bayesian Analysis of Two-Stage Randomized Experiments in the Presence of Interference, Treatment Nonadherence, and Missing Outcomes

John Morgan - Virginia Tech
A New Look at the Search Design Concept

Simon Mak - Duke University
Design and analysis of multi-stage multi-fidelity computer experiments, with application to emulation of heavy-ion collisions

Xinwei Deng - Virginia Tech
A Machine Learning Perspective for Experimental Design via Tight Mutual Information

3:50 – 5:50 **October 7 - Parallel Session 2B: Dail Room EUC**
Computational Advertising
Chair: David Banks – Duke University
david.banks@duke.edu

Patrick LeBlanc - Duke University
Cross-Domain Recommender Systems

Jennie Huang - Duke University
Online Controlled Experiments: Top Challenges and Solutions

Mingyan Li - UNC Greensboro
Probabilistic Factorization Matrix

Ryan Tang - Duke University
Ad Marketplace Optimization Towards Auto-Bidding

3:50 – 5:50 **October 7 - Parallel Session 2C: Claxton Room EUC**
Modern Models and Computational Methods for Today's Complex Data
Chair: Anuradha Roy, University of Texas San Antonio
Anuradha.Roy@utsa.edu

Emily Hector – North Carolina State University
Mean Structure Learning with High-Dimensional Correlated Data

Sean L. Simpson – Wake Forest University School of Medicine
Analytical Tools for Whole-Brain Networks: Fusing Statistics and Network Science to Understand Brain Function

Ben Seiyon Lee - George Mason University
A Scalable Partitioned Approach to Model Massive Nonstationary Non-Gaussian Spatial Datasets

Marco Ferreira – Virginia Tech
Bayesian Analysis of GLMMs with Nonlocal Priors for Genome-Wide Association Studies

Xiaoyan (Iris) Lin – University of South Carolina
Bayesian Gaussian Copula Graphical Model for Ordinal Data

3:50 – 5:50 **October 7 - Parallel Session 2D: Alexander Room EUC**
Inference in Higher-Order Markov Models
Chair: Donald E.K. Martin - North Carolina State University
demarti4@ncsu.edu

Soumendra Lahiri – North Carolina State University
A Scalable Method for Fitting Sparse Markov Models

Tuhin Majumder - Duke University
Fitting Sparse Markov Models Through Regularization

Younghoon Kim – UNC Chapel Hill
Latent Gaussian Dynamic Factor Modeling and Forecasting for Multivariate Count Time Series

Donald E.K. Martin - North Carolina State University
Inference for Hidden Sparse Markov Models

3:50 – 5:50 **October 7 - Parallel Session 2E: Kirkland Room EUC**
Sample Survey Methodology
Chair: Javid Shabbir – University of Wah, Pakistan
javidshabbir@gmail.com

Zaheen Khan – Federal Urdu University of Arts, Science and Technology, Islamabad
Application of Modified Systematic Sampling in Auto-correlated Populations

Hina Khan – GC College University, Pakistan
On Estimation and Monitoring of Population Mean Using Generalized Neutrosophic Ratio-Type Exponential Estimator Under Neutrosophic Exponentially Weighted Moving Average Scheme

Zaheen Khan – Federal Urdu University of Arts, Science and Technology, Islamabad
Unbiased Estimation of Variance of Sample Mean in Systematic Sampling

Javid Shabbir - University of Wah, Pakistan
Performance of Optional Unrelated Question Randomized Response Models under Two-Stage Stratified Cluster Sampling for Estimation of Population Proportion and Sensitivity Level

6:00 – 7:20 **Statistical Consulting in a University Setting: Modern Challenges and Enduring Issues: EUC Auditorium**
Chair: Daniel Hall – University of Georgia
danhall@uga.edu

Panelists:
Daniel Hall – Director of University of Georgia Statistical Consulting Session

Pat Gerard, Director of the Statistics and Mathematics Consulting Center at Clemson University

Krista Gile, Co-Director of Statistical Consulting & Collaboration Services at the University of Massachusetts-Amherst

Marianne Huebner, Director of the Center for Statistical Training & Consulting at Michigan State University

7:30 – 9:00 **Reception: Cone Ballrooms, EUC**
Hosted by: Springer

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

October 8, 2022, Saturday

7:45 & 8:15 **Shuttles from Holiday Inn to UNCG**
Please do not wait for the last shuttle. You may remain stranded and may have to take an Uber.

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

9:15 – 10:40 **Keynote Address II & Plenary Talk 4 EUC Auditorium**
Chair: John Stufken – George Mason University
jstufken@gmu.edu

9:15 – 10:00 **Keynote Address II**
Srinivas Varadhan – Courant Institute/NYU
The Polaron Problem

10:00 – 10:40 **Plenary Talk 4**
Richard Davis - Columbia University
Time Series Estimation of the Dynamic Effects of Disaster-Type Shocks

10:40 – 11:00 **Coffee Break**

11:00 – 1:00 **October 8 - Parallel Session 3A: Sharpe EUC**
Design of Experiments 2: Subdata Selection
Chair: John Morgan – Virginia Tech
jpmorgan@vt.edu

HaiYing Wang - University of Connecticut
Maximum Sampled Conditional Likelihood for Informative Subsampling

John Stufken - George Mason University
Subdata Selection With a Large Number of Variables

Lin Wang - Purdue University
Balanced Subsampling for Big Data with Categorical Predictors

Min Yang - University of Illinois at Chicago
Selecting Nearly Optimal Subdata

11:00 – 1:00 **October 8 - Parallel Session 3B: Dail Room EUC**
Object Data Analysis
Chair: Victor Patrangenaru – Florida State University
vic@stat.fsu.edu

Ian Dryden – Florida International University
Shape Analysis of Molecular Dynamics Data

Adam Dixon – Florida State University
Investigating Two Possible Origins of SARS-CoV-2: An RNA Analysis on Tree Spaces

Aaid Algahtani - Florida State University
Limit Theorems for Object Data with Applications to 3D Image Data Analysis

Seungee Choi – Florida State University
Estimation of Spherical Depth on Object Spaces

Vic Patrangenaru - Florida State University
RCD and TDA for 2D Scenes Extracted From Electronic Images

11:00 – 1:00 **October 8 - Parallel Session 3C: Claxton Room EUC**
Topological Data Analysis in Machine Learning and Statistics - I
Organizers: Thomas Weighill – UNC Greensboro, Tom Needham - Florida State University
Chair: Thomas Weighill – UNC Greensboro
t_weighill@uncg.edu

Radmila Sazdanovic – North Carolina State University
Mapper-Type Algorithms: Extensions and Generalizations

Woojin Kim – Duke University
Computing Generalized Rank Invariant via Zigzag Persistence

Vasileios Maroulas - University of Tennessee, Knoxville
Random Persistence Diagram Generator

Tom Needham - Florida State University
Hypergraph Co-Optimal Transport

11:00 – 1:00 **October 8 - Parallel Session 3D: Alexander Room EUC**
Statistical Applications in Health Sciences
Chair: Rishi Chakraborty – Duke University
rishi.c@duke.edu

MD Shahjahan - Daffodil International University, Bangladesh
Predictors of Depression and Anxiety Among Urban Population During COVID-19: An Online Cross-Sectional Survey

Sophia Waymyers - Francis Marion University
Modeling Negatively Skewed Survival Data in Accelerated Failure Time Models

Shah Golam Nabi - DGHS, Ministry of Health and Family Welfare of Bangladesh
Initiatives of Emergency Response and COVID-19 Pandemic Preparedness for Health System Strengthening to Combat COVID -19 Pandemic in Bangladesh

Mohammad Saiful Islam – DGHS, Ministry of Health and Family Welfare of Bangladesh
Impact of Establishment of Liquid Medical Oxygen System Indifferent Health Facility to Provide Quality Care of Critical Patient During Covid-19 Crisis in Bangladesh

Abu Minhajuddin – University of Texas Southwestern
Identifying Subgroups of Adolescents with Depression Suicidal Ideation: A Look at the TX-YDSRN Data

11:00 – 1:00 **October 8 - Parallel Session 3E: Kirkland Room EUC**
Undergraduate Research
Chairs: Steven Miller (Williams College) and Lazaros Gallos (Rutgers University)
sjm1@williams.edu, lgallos@dimacs.rutgers.edu

Anthony Lee - Milton Academy
Generalizing the German Tank Problem

Flora Shi – Duke University
ESPs: A New Cost-Efficient Sampler for Expensive Posterior Distributions

Jenny Huang – Duke University
Online Controlled Experiments: Top Challenges and Solutions

Katie Anderson - Brigham Young University-Idaho
A New Benford Test for Clustered Data with Applications to American Elections

Rhys O'Higgins - Macalester College
Subdata Selection and TreeS

11:00 – 1:00 **October 8 - Parallel Session 3F: Dogwood Room EUC**
Functional Data Analysis: New Directions and Innovations
Organizer: Lily Wang – George Mason University
Chair: Xinyi Li – Clemson University
lixinyi@clemson.edu

Xinyi Li – Clemson University
Individualized Treatment Regimes Incorporating Imaging Features

Wanli Qiao – George Mason University
Embedding Functional Data: Multidimensional Scaling and Manifold Learning

Guannan Wang – College of William and Mary
Statistical Inference for Mean Functions of 3D Functional Objects

Guanqun Cao – Auburn Uniersity
Deep Neural Network Classifier for Functional Data

1:00 – 2:00 **Lunch at Cone Ballrooms, EUC**

2:00 – 3:30 **October 8 - Plenary Talks 5/6: EUC Auditorium**
Chair: Sat Gupta – UNC Greensboro
sngupta@uncg.edu

2:00 – 2:40 **Karen Kafadar (ASA President 2019) - University of Virginia**
To Screen or Not to Screen? Using Data From Randomized Screening Trials to Quantify Risks & Benefits of Cancer Screening

2:45 – 3:25 **Eiren Jacobson - University of St. Andrews, UK**
State-Space Models for Marine Mammal Populations

3:30 – 3:50 **Coffee Break**

3:50 – 5:50 **October 8 - Parallel Session 4A: Sharpe EUC**
Design of Experiments 3: Screening Experiments
Chair: Abhyuday Mandal – University of Georgia
amandal@stat.uga.edu

David Edwards - Virginia Commonwealth University
Strategies for Supersaturated Screening: Group Orthogonal and Constrained Var(s) Designs

Qian Xiao - University of Georgia
Maximum One-Factor-At-A-Time Designs for Screening in Computer Experiments

Rakhi Singh - Binghamton University
Design Selection for Supersaturated Designs

Ryan Lekivetz – SAS
A Family of Orthogonal Main Effects Screening Designs for Mixed Level Factors

3:50 – 5:50 **October 8 - Parallel Session 4B: Dail Room EUC**
Undergraduate Research

Best Presentation Award Session for Undergraduate Students
Chairs: Steven Miller (Williams College) and Lazaros Gallos (Rutgers University)
sjm1@williams.edu, lgallos@dimacs.rutgers.edu

Yiren Hou - University of Georgia
Optimal Design for Ordinal Categorical Regression on Milk Fiber Strength

Joia Zhang - University of Washington, Seattle
An Optional Quantitative Mixture RRT Model that Accounts for Lack of Trust

Bailey Meche – University of Louisiana Lafayette
DiseaseNet: a Unified Approach to Disease Detection

Zoe McDonald (Boston University) and Livia Betti (University of Rochester)
Benfordness of Measurements Resulting From Box Fragmentation

William McCance - UC Santa Barbara
Binary Randomized Response Technique (RRT) Models Under Measurement Error

3:50 – 5:50 **October 8 - Parallel Session 4C: Claxton Room EUC**
Topological Data Analysis in Machine Learning and Statistics - II
Organizers: Thomas Weighill – UNC Greensboro, Tom Needham - Florida State University
Chair: Tom Needham – Florida State University
TNEEDHAM@fsu.edu

Jakini Kauba - Clemson University
An Analysis of Demographic Trends Using Topological Data Analysis

Neil Pritchard – UNC Greensboro
Coarse Embeddability of the Space of Persistence Diagrams and Wasserstein Space

Yuan Wang - University of South Carolina
Topological Inference on Brain Signals

Don Sheehy – North Carolina State University
Semi-Supervised TDA

3:50 – 5:50 **October 8 - Parallel Session 4D: Alexander Room EUC**
Statistical Ecology
Chair: Eiren Jacobson, University of St. Andrews, UK
eiren.jacobson@st-andrews.ac.uk

Frances Buderman – Penn State University
A Life-History Spectrum of Population Responses to Simultaneous Change in Climate and Land Use

Leslie New - Ursinus College
Balancing Wind Energy Production and Bat Fatalities

Eiren Jacobson, University of St. Andrews, UK
Quantifying The Response Of Blainville's Beakedwhales to U.S. Naval Sonar Exercises in Hawaii

3:50 – 5:50 **October 8 - Parallel Session 4E: Kirkland EUC**
Recent Advances in Statistical Estimation and Testing
Chair: Qin Shao, University of Toledo
Qin.Shao@utoledo.edu

Jing Wang, University of Illinois-Chicago
Semiparametric Estimation of Non-Ignorable Missingness with Refreshment Sample

Jiancheng Jiang, University of North Carolina at Charlotte
Testing for Relevance in Partially Parametric Models with Parametric Nulls

Hongbin Zhang, University of Kentucky
Statistical Methods For Interval-Censored Multistate Data And Mis-Measured Covariates With Application In HIV Care

Zhezhen Jin, Columbia University
A Step-Wise Multiple Testing With Linear Regression Models

Anuradha Roy – University of Texas at San Antonio
Linear Models for Doubly Multivariate data with Exchangeably Distributed Errors

3:50 – 5:50 **October 8 - Parallel Session 4F: Dogwood Room EUC**
Bayesian Methods for Physical Science and Engineering Applications
Chair: Simon Mak – Duke University
sm769@duke.edu

Irene Ji – Duke University
A Graphical Multi-Fidelity Gaussian Process Model, With Application to Emulation of Heavy-Ion Collisions

Xiaojun Zheng – Duke University
PERCEPT: a New Online Change-Point Detection Method Using Topological Data Analysis

Kevin Li – Duke Uniersity
Variational Manifold-Embedded Gaussian Process Modeling, With Applications to Aircraft Engine Design

Didong Li – UNC Chapel Hill
Probabilistic Contrastive Principal Component Analysis

Ye Tian – Columbia University
Unsupervised Multi-task and Transfer Learning on Gaussian Mixture Models

6:00 – 6:45 **JSTP Editorial Board Meeting – Azalea Room, EUC**

6:45 – 9:00 **Conference Banquet & Student Awards – Cone Ballrooms, EUC**

9:00 & 9:30 **Shuttles back to Holiday Inn**
Departures from outside of EUC on Stirling Street: 9:10 pm, 9:30 pm
Please do not wait for the last shuttle. You may remain stranded and may have to take an Uber

October 9, 2022, Sunday

9:00 **Shuttle from Holiday Inn to UNCG**
Just one shuttle today

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

10:00 – 12:00 **October 9 - Parallel Session 5A: Sharpe EUC**
Design of Experiments 4: Computer Experiments
Chair: John Stufken – George Masson University
jstufken@gmu.edu

Abhyuday Mandal - University of Georgia
Modeling and Active Learning for Experiments with Quantitative-Sequence Factors

Chenlu Shi - Colorado State University
A Projection Space-Filling Criterion and Related Optimality Results

Lulu Kang - Illinois Institute of Technology
Dimension Reduction for Gaussian Process Models via Convex Combination of Kernels

Chih-Li Sung - Michigan State University
Functional-Input Gaussian Processes with Applications to Inverse Scattering Problems

10:00 – 12:00 **October 9 - Parallel Session 5B: Dail Room EUC**
Statistical Applications in Health Sciences and Nursing Research
Co-Chairs: Susan Letvak and Thomas P. McCoy - UNC Greensboro
tpmccoy@uncg.edu, saletvak@uncg.edu

Crystal Modde Epstein and Thomas P. McCoy - UNC Greensboro
Mixed-effects Cosinor Analysis of Cortisol Patterns Among Pregnant Women

Forgive Avorgbedor – UNC Greensboro
Effects of Neighborhood and Household Socioeconomic Disadvantage on Postpartum Weight Retention

April W. Messer – Western Carolina University
Nurses' Experiences Caring for Patients with Opioid Use Disorders

Thomas P. McCoy (UNC Greensboro) and Marjorie Jenkins (Cone Health)
Time-to-ICU Transfer by Frailty Severity using Electronic Health Record Data: Does Nursing Flowsheet Data Matter?

10:00 – 12:00 **October 9 - Parallel Session 5C: Claxton Room EUC**
Randomized Response Models
Chair: Sadia Khalil – UNC Greensboro & Lahore College for Women University
sadia_khalil@hotmail.com

Badr Aloraini – Wake Forest University
Estimation of Population Variance for a Sensitive Variable in Stratified Sampling Using Randomized Response Technique

Pujita Sapra – UNC Greensboro
Accounting for Lack of Trust in Optional Binary RRT Models Using a Unified Measure of Privacy and Efficiency

Wendy Shou – UNC Greensboro
Kernel Density Estimation Using Additive Randomized Response Technique (RRT) Models

Sadia Khalil – UNC Greensboro & Lahore College for Women University
Mean Estimation Using RRT Models Under Two-Phase Simple & Stratified Random Sampling Designs

10:00 – 12:00 **October 9 - Parallel Session 5D: Dogwood EUC**
Applied Statistics
Chair: Suresh Chander Malik – M D University Rohtak
sc_malik@rediffmail.com

Suresh Chander Malik – M D University Rohtak, India
On Use of Regression Approach for Reliability Variation of a Parallel-Series System

Fatema Ruhi – Southeast Missouri State University
A Comparison of Multiple Testing Procedures for Controlling the False Discovery Rate Under Unequal Variances

Sheela Kumari – BPSIHL, BPSMV, Khanpur, Sonepat, India
Development of the Subject Statistics: A Historical Perspective

10:00 – 12:00 **October 9 - Parallel Session 5E: Alexander Room EUC**
Statistical Modeling
Chair: Haimeng Zhang – UNC Greensboro
h_zhang5@uncg.edu

Sarangan Balasubramaniam – University of Georgia
Spatial Prediction for Axially Symmetric Process on Spheres

Xiaohuan (Max) Xue – UNC Greensboro
Constructing Covariance Functions for Axially Symmetric Processes on the Sphere

Yuyan Yi – Auburn University
CW_ICA: An Efficient Dimensionality Selection Method for Independent Component Analysis

Haolei Weng - Michigan State University
Signal-To-Noise Ratio Aware Minimaxity for Sparse Gaussian Sequence Models

10:00 – 12:00 **October 9 - Parallel Session 5F: Kirkland Room EUC**
Some Topics on Robust Machine Learning
Chair: Xiaoli Gao – UNC Greensboro
x_gao2@uncg.edu

Quefeng Li - UNC Chapel Hill
Decomposition of Variation of Mixed Variables by a Latent Mixed Gaussian Copula Model

Wei Ning - Bowling Green State University
Monitoring Sequential Structural Changes in Penalized High-Dimensional Linear Models

Yao Li – UNC Chapel Hill
Trusted Aggregation (TAG): Model Filtering Backdoor Defense in Federated Learning

Matt Jester – UNC Greensboro
Robust Topological Classification with Applications in Firn Data Analysis

Zhengwu Zhang – UNC Chapel Hill
High-Dimensional Spatial Quantile Function-on-Scalar Regression

12:10 **Closing Remarks, EUC Auditorium**

12:30 – **Lunch at Cone Ballrooms, EUC**

1:45 – **Shuttle back to Holiday Inn**
Departures from outside of EUC on Stirling Street: 2:00 pm

Abstracts of the talks

Statistical Shape Analysis of 3DFN Data: Analysis via Deformetrica
Mithun Acharjee
University of Alabama Birmingham
Coauthor: Matthew Reimherr, Pennsylvania State University

Two important aspects of shape analysis are to estimate the mean shape from a given set of shapes and to estimate a shape trajectory as close as to the observed shapes to determine the continuous evaluation of shapes over time. A Deterministic Atlas (DA) model is used to compute the mean shape from a set of shapes which builds a generalization of a typical representation by preserving the characteristics of the original shapes. The Geodesic Regres- sion (GR) is used to estimate the continuous shape evaluation at a certain time within its intervals where the mean face obtained from the DA model can be used as a baseline shape to initiate the program. This research work introduces the application of the DA model and GR model using Deformetrica (shape analysis software) where the data consideration was the three-Dimensional Facial Norm (3DFN) database deposited in the FaceBase consortium. Deformetrica executes the deterministic atlas model to estimate the mean 3-dimensional fa- cial object. This estimated mean face is used as an initial template shape on GR which provides an estimated shape trajectory of the facial objects. This geodesic shape trajectory is a geodesic flow of diffeomorphisms acting on the above baseline template shape to estimate the continuous 3D facial evaluation with age varying continuously within its range. This research also describes the technical details of using Deformetrica in a high-performance computing environment while dealing with a 3D geometric object with a high number of landmark points.

Limit Theorems for Object Data with Applications to 3D Image Data Analysis
Aaid Algahtani
Florida State University
ama17s@fsu.edu
Coauthor: Vic Patrangenaru

In recent years imaging data has become the most overwhelming type of data, due availability of smart phones to the public. This article uses statistical shape analysis to analyze such data. One develops nonparametric procedures for comparing two extrinsic means on the oriented projective shape space $\overrightarrow{P}(\mathbb{R}^{m+1})$ of k-ads in general position in $\mathbb{R}^{m+1}$. Applications to 3D bioshape analysis for data extracted from digital camera images are given. For small samples, the tests are run based on Efron's nonparametric bootstrap.

Estimation of Population Variance for a Sensitive Variable in Stratified Sampling Using Randomized Response Technique
Badr Aloraini
Wake Forest University
aloraib@wfu.edu
Coauthors: Sadia Khalil, Lahore College for Women University; Muhammad Nouman Qureshi, National College of Business Administration and Economics; Sat Gupta, University of North Carolina at Greensboro

In this paper, Randomized Response technique (RRT) is used to propose some separate and combined variance estimators for a sensitive variable using stratified random sampling. The performances of the proposed estimators are examined using a unified measure of respondent privacy and estimator efficiency.

A New Benford Test for Clustered Data with Applications to American Elections

Katie Anderson
Brigham Young University-Idaho
andersonkatie1998@gmail.com
Coauthors: Kevin Dayaratna, Center for Data Analysis, The Heritage Foundation; Drew Gonshorowski, Center for Data Analysis, The Heritage Foundation; Steven J. Miller, Williams College

Benford's law describes the distribution of leading digits that occur in many real-life datasets, or in other words, the probability that a leading digit d will occur. In some situations, if the leading digits do not follow Benford's law when they are expected to, it can be a sign of malfeasance. A problem with classic first digit applications of Benford's law is the law's inapplicability to clustered data. This has especially been a common problem with election data. We discuss a new proposal that will allow Benford's first digit analysis to be useful when working with clustered data. Namely, performing a first digit analysis after converting the data to base 3 (1-BL 3). This conversion to base 3 will spread out the data, thus rendering Benford's law significantly more useful. We present 1-BL 3 as a useful method when applied to election data. We test the efficacy of our approach on synthetic election data using discrete Weibull modeling, finding in many cases that election data often conforms to 1-BL 3. We then apply 1-BL 3 analysis to selected states from the 2004 US Presidential election to detect potential statistical anomalies. We also find that in some cases there is data that is flagged as suspicious that may actually not be. This is a joint work with Kevin Dayaratna, Drew Gonshorowski, and Steven J. Miller.

Effects of Neighborhood and Household Socioeconomic Disadvantage on Postpartum Weight Retention

Forgive Avorgbedor
Department of Adult Health Nursing, School of Nursing, UNC Greensboro
f_avorgbedo@uncg.edu
Coauthors: Thomas P. McCoy, Laurie Wideman, Lenka H. Shriver, Cheryl Buehler, Esther M. Leerkes, UNC Greensboro

This study investigated the pathways by which race, neighborhood socioeconomic disadvantage (NSD), and household socioeconomic disadvantage (HSD) predict subsequent maternal postpartum weight retention (PWR). One hundred seventy-six (n = 176) racially diverse women were studied from the 3rd trimester to 6 months postpartum. NSD was defined by information from the American Community Survey based on women's census tract and self-reports of neighborhood healthy food availability, safety, violence, and walking environment. HSD included food insecurity, income-to-needs ratio, and maternal education. PWR was operationalized as a 6-month postpartum weight minus pre-pregnancy weight. Data were analyzed using structural equation modeling with bootstrapped confidence intervals to estimate indirect effects. It was concluded that to prevent PWR, education on behavior change to lose weight is essential, and it must be offered in the context of basic resources, at both the neighborhood and household levels.

Spatial Prediction for Axially Symmetric Process on Spheres

Sarangan Balasubramaniam
University of Georgia
s_balasu@uncg.edu
Coauthor: Haimeng Zhang, UNC Greensboro

Spatial prediction, or so-called kriging, is one of the ultimate goals in spatial data analysis. The basic idea of kriging is to use the values of a geographic variable at some locations to estimate the value(s) that are unknown at other locations. In this dissertation, we consider the spatial prediction when a random process is axially symmetric on the sphere. More specifically, we first decompose an axially symmetric process as Fourier series on circles, where the Fourier random coefficients can be expressed as circularly-symmetric complex

random processes. The estimation of covariance functions for complex random processes is then obtained through both parametric and non-parametric approaches, where least squared error estimation and the Wavelet-Galerkin methods are applied, respectively. Ordinary kriging is then conducted on possibly complex random fields and predicted data values are computed through the inverse Discrete Fourier transformation. All the above approaches and results are demonstrated through simulation studies.

Route Choice Under Uncertainty: An Adversarial Risk Analysis

David Banks
Duke University
banks@stat.duke.edu

Problems in counterterrorism and corporate competition have prompted research that attempts to combine statistical risk analysis with game theory in ways that support practical decision making. This article applies these methods of adversarial risk analysis to the problem of selecting a route through a network in which an opponent chooses vertices for ambush. The motivating application is convoy routing across a road network when there may be improvised explosive devices and imperfect intelligence about their locations.

A Life-History Spectrum of Population Responses to Simultaneous Change in Climate and Land Use

Frances Buderman
Ecosystem Science and Management, Pennsylvania State University
fbuderman@psu.edu
Coauthors: James H. Devries, Institute for Wetland and Waterfowl Research, Ducks Unlimited Canada; David N. Koons, Fish, Wildlife, and Conservation Biology, Colorado State University"

Climate and land-use change are two of the primary threats to global biodiversity; however, each species within a community may respond differently to these facets of global change. Although it is typically assumed that species use the habitat that is advantageous for survival and reproduction, anthropogenic changes to the environment can create ecological traps, making it critical to assess both habitat selection (e.g., where species congregate on the landscape) and the influence of selected habitats on the demographic processes that govern population dynamics (e.g., survival and reproduction). We hypothesized that species-specific responses to environmental change would scale with life-history traits, specifically: maternal investment, nesting phenology, and female breeding site fidelity. We used a long-term (1958-2011), large-scale, multi-species data set for waterfowl that spans the United States and Canada to estimate species-specific responses to climate and land cover variables in a landscape that has undergone significant environmental change across space and time. We first estimated the effects of change in climate and land cover variables on habitat selection and population dynamics for nine species using a hierarchical Bayesian model that separates the two processes. We then used multiple imputation and a Bayesian model that accounted for phylogenetic relatedness to identify relationships between life history traits (nesting phenology, maternal investment, and breeding site fidelity) and responses to climate and land use changes. We detected several significant relationships between life-history traits, particularly nesting phenology, and species' responses to environmental change. One species, the early-nesting northern pintail, was consistently at the extreme end of responses to land cover and climate predictors and has been a species of conservation concern since their population began to decline in the 1980s. They, and the blue-winged teal, also demonstrated a positive habitat selection response to the proportion of cropland on the landscape that simultaneously reduced abundance the following year, indicative of susceptibility to ecological traps. By distilling the diversity of species' responses to environmental change within a community, our methodological approach and findings will help improve predictions of community responses to global change and can inform multi-species management and conservation plans in dynamic landscapes that are based on simple tenets of life-history theory.

Deep Neural Network Classifier for Functional Data
Guanqun Cao
Auburn University
gzc0009@auburn.edu
Coauthors: Shuoyang Wang, Yale University; Zuofeng Shang, New Jersey Institute of Technology

We propose a new approach, called as functional deep neural network (FDNN), for classifying functional data. Specifically, a deep neural network is trained based on the principal components of the training data which shall be used to predict the class label of a future data function. Unlike the popular functional discriminant analysis approaches which rely on Gaussian assumption, the proposed FDNN approach applies to general non-Gaussian multi-dimensional functional data. Moreover, when the log density ratio possesses a locally connected functional modular structure, we show that FDNN achieves minimax optimality. The superiority of our approach is demonstrated through both simulated and real-world datasets.

Comparison of Various Algorithms in Optimal Obsta Placement with Disambiguation (OPD) Problem
Elvan Ceyhan
Auburn University
ceyhan@auburn.edu
Coauthor: Polat Charyev, MAP Akademi, Istanbul, Turkey

In the optimal obstacle placement with disambiguation (OPD) problem, we consider various obstacle placement schemes against different clutter point realizations sampled from Poisson, clustered, and regular spatial point patterns. Obstacles are placed according to a Poisson distribution in various window types such as linear strips, V-shaped and W-shaped regions. Based on this construction, reset disambiguation (RD) algorithm is used to determine the traversal length with the disambiguation option. Additionally, we also implement the M4, M8, M16 algorithms, which are mainly based on the effective choice of a subset of possible disambiguations to determine the traversal length. We compare the empirical performances of RD, M4, M8 and M16 algorithms based on extensive Monte Carlo simulations.

Estimation of Spherical Depth on Object Spaces
Seunghee Choi
Florida State University
sc17x@fsu.edu
Coauthor: Victor Patrangenaru

The depth has been extensively studied in nonparametric statistics. Various forms of depth have been proposed to provide a center-outward ranking, for example, location depth by Tukey, simplicial depth by Liu, projection depth by Zuo and Serfling, and spherical depth by Elmore and Fraiman. In this project, we employ the classical spherical depth and extend it to general random objects. The uniform consistency and limiting distribution of an empirical spherical depth function on an object space are studied. We illustrate our approach in simulated data and real data examples.

A Brief Parametric Analysis of Catastrophic or Disastrous Hurricanes That Have Hit the Florida Keys between 1900 and 2000

Joy D'Andrea
University of South Florida
jdandrea@mail.usf.edu

The most intense and catastrophic hurricanes on record to hit the Florida Keys during 1900 to 1950 were in 1919, and 1935. From 1950 to 2000, the most intense hurricanes to hit or affect the Florida Keys were in 1960, 1965, and 1992. In this talk, we will present a brief parametric analysis of the hurricanes that have hit the Florida Keys in the last 100 years. This analysis will include the descriptive statistics, best fit probability distribution of the latitude of the catastrophic hurricanes and a confidence interval that detects the average latitude of hurricanes (category 3 or higher) which have hit the Florida Keys in the last 100 years.

Air Quality and Lung Cancer: Analysis via Local Control

Kumer Das
University of Louisiana at Lafayette
kumer.das@louisiana.edu
Coauthors: Mithun Acharjee, University of Alabama Birmingham; S. Stanley Young

There are literature claims of a concentration-response, C-R, link between air quality, as measured by particulate matter, and lung cancer. There is a need to understand if this possible link is influenced by other variables. Our first idea is to start with Environmental Quality Index, EQI, variables as covariates and explore possible C-R het-erogeneity using Local Control (difference and regression) analysis, LCD and LCR. The first step of LC is to clus-ter the units, US counties so that counties with similar covariates are together and hence covariate controlled. Our second idea is to use causal analysis methods to better understand the links between the covariates, air quality, and lung cancer. We add percent smoking as an additional covariate into the causal analysis. We note possible prob-lems with the construction and use of the EQIs. We discover several things. Overall, there is positive association between lung cancer mortality and PM2.5. Both LCD and LCR slope show a positive effect. We note the superiori-ty of the LCR over LTD based on higher R-squared value and lower root mean squared error. Using recursive par-titioning we note considerable heterogeneity in the C-R relationship. EQI variables do not capture smoking, so we add smoking and addition direct variables to our causal analysis. Using causal analysis methods, primarily partial correlations, we find complex relationships among the variables, which we display with causal diagrams. These diagrams do not link lung cancer mortality to particulate matter.

Keywords: PM2.5, Lung cancer mortality, Local control analysis, Local treatment difference, Local control regression

Time Series Estimation of the Dynamic Effects of Disaster-Type Shocks

Richard Davis
Columbia University
davis.richarda@gmail.com
Coauthor: Serena Ng

This paper provides three results for SVARs under the assumption that the primitive shocks are mutually independent. First, a framework is proposed to accommodate a disaster-type variable with infinite variance into a SVAR. We show that the least squares estimates of the SVAR are consistent but have non-standard asymptotics. Second, the disaster shock is identified as the component with the largest kurtosis. An estimator that is robust to infinite variance is used to recover the mutually independent components. Third, an independence test on the residuals pre-whitened by the Choleski decomposition is proposed to test the restrictions imposed on a SVAR. The test can be applied whether the data have fat or thin tails, and to over as well as exactly identified models. Three applications are considered. In the first, the independence test

is used to shed light on the conflicting evidence regarding the role of uncertainty in economic fluctuations. In the second, disaster shocks are shown to have short term economic impact arising mostly from feedback dynamics. The third uses the framework to study the dynamic effects of economic shocks post-covid. (This is joint work with Serena Ng.)

A Machine Learning Perspective for Experimental Design via Tight Mutual Information

Xinwei Deng
Department of Statistics, Virginia Tech
xdeng@vt.edu
Coauthor: Qing Guo

Effective data collection is very important in data science. In this work, we present a machine learning perspective for experimental design. The proposed work is framed under the Bayesian optimal experimental design (BOED) setup, where the mutual information (MI) between data and model parameters is maximized via optimizing the experimental parameters. However, directly applying existing MI estimators may not work since they rely on explicit knowledge of the model likelihood. To overcome these limitations, we revisit the popular variational MI bounds from the lens of unnormalized statistical modeling and convex optimization. Our investigation leads to a novel, simple, and powerful contrastive MI estimator for experimental design. The performance of the proposed method is evaluated by both the likelihood-free experimental design and the amortized sequential experimental design.

Investigating Two Possible Origins of SARS-CoV-2: An RNA Analysis on Tree Spaces

Adam Dixon
Florida State University
ad18g@fsu.edu
Coauthors: Roland Moore, Victor Patrangenaru

Using the construction of Billera, Holmes, and Vogtmann, rooted phylogenetic trees with three leaves (3-trees) are regarded as a two-dimensional stratified space. Since 3-trees include exactly one interior node and one interior edge, branch lengths corresponding to the root-to-interior-node edge and the interior edge are used to define an evolutionary distance between the root and the most recent common ancestor (MRCA) of the leaves. With this definition, the proximities of various roots to the leaves can be compared. To apply this method, recent RNA sequences of SARS-CoV-2 collected from multiple sources are aligned and used to compare sets of 3-trees with two different roots: bat coronavirus RatG13 (BatCov RatG13) and a SARS-CoV-2 sequence taken from a human subject at the onset of the COVID-19 pandemic. Nonparametric bootstrap methods are used to infer the difference in mean evolutionary distance between the two root groups, and the early-sequenced SARS-CoV-2 root appears to be evolutionarily closer to the MRCA than BatCov RaTG13.

Bayesian Multi-Study Clustering of Sepsis Patients: Utilizing Prior Information and Testing Cluster Discovery

Alexander Dombowsky
Duke University
alexander.dombowsky@duke.edu
Coauthors: Amy Herring, David Dunson

Identifying subtypes of life-threatening but broadly defined conditions such as sepsis is critical for designing effective therapies. Clinical data–consisting of readily measurable patient signs upon presentation to hospital–have been clustered in past studies for determining sepsis subtypes, often using algorithmic clustering methods such as k-means. In this talk, we cluster a new cohort of sepsis patients with a small sample size and vastly different demographic information than previous cohorts, leveraging available information by fitting a Bayesian Gaussian mixture model with highly informative priors. Our model also accounts for the possibility of discovering a cluster in our cohort that was not present in previous cohorts due to demographic differences, and we calculate the Watanabe–Akaike information criterion (WAIC) for several candidate models to validate these newly discovered clusters. We also compare our clusters to those derived from other algorithms, and explain how incorporating prior information leads to interpretable and medically consistent clusters.

Shape Analysis of Molecular Dynamics Data

Ian Dryden
Florida International University
idryden@fiu.edu

Molecular dynamics simulations produce huge datasets of temporal sequences of molecules. It is of interest to summarize the shape evolution of the molecules in a succinct, low-dimensional representation. However, Euclidean techniques such as principal components analysis (PCA) can be problematic as the data may lie far from in a flat manifold. Principal nested spheres gives a fundamentally different decomposition of data from the usual Euclidean subspace based PCA. Subspaces of successively lower dimension are fitted to the data in a backwards manner with the aim of retaining signal and dispensing with noise at each stage. We adapt the methodology to 3D subshape spaces and provide some practical fitting algorithms. The methodology is applied to cluster analysis of peptides, where different states of the molecules can be identified. Also, the temporal transitions between cluster states are explored. Further molecular modelling tasks include resolution matching, where coarse resolution models are backmapped into high resolution (atomistic) structures.

Strategies for Supersaturated Screening: Group Orthogonal and Constrained Var(s) Designs

David Edwards
Virginia Commonwealth University
dedwards7@vcu.edu
Coauthors: Maria Weese, Miami University; Jonathan Stallrich, North Carolina State University; Byran Smucker, Miami University

Despite the vast amount of literature on supersaturated designs (SSDs), there is a scant record of their use in practice. We contend this imbalance is due to conflicting recommendations regarding SSD use in the literature as well as the designs' inabilities to meet practitioners' analysis expectations. To address these issues, we first summarize practitioner concerns and expectations of SSDs as determined via an informal questionnaire. Next, we discuss and compare two recent SSDs that pair a design construction method with a particular analysis method. The choice of a design/analysis pairing is shown to depend on the screening objective. Group orthogonal SSDs, when paired with our new, modified analysis, are demonstrated to have high power even with many active factors. Constrained positive Var(s)-optimal designs, when paired with the

Dantzig selector, are recommended when effect directions can be credibly specified in advance; this strategy reasonably controls Type I error rates while still identifying a high proportion of active factors.

Mixed-Effects Cosinor Analysis of Cortisol Patterns Among Pregnant Women

Crystal Modde Epstein and Thomas P. McCoy
Department of Family & Community Nursing, School of Nursing, UNC Greensboro
cmepstein@uncg.edu

Chronobiological methods to study circadian rhythmicity has made many methodological advances in the last thirty years. Among them is use of cosinor analysis to study daily cycles of repeatedly measured biophysiological phenomena such as blood pressure, heart rate variability, and salivary cortisol. This study extends the recently developed mixed-effects cosinor modeling based on Hou et al. (2021) in the cosinoRmixed R package to include multiple continuous covariates of age and structural racism measured by the Index of Concentration at the Extremes using SAS PROC NLMIXED in modeling cortisol patterns among a sample of pregnant women. Nonlinear cosinor-derived parameter estimates are discussed, and future directions of the methodology will be delineated.

Two-Sample Testing With Local Community Depth

Ciaran Evans
Wake Forest University
evansc@wfu.edu
Coauthor: Kenneth Berenhaut

In this talk, we develop a nonparametric two-sample hypothesis test based on local community depth, a new data depth calculated from pairwise distances between observations. Unlike many common "center-outwards"' measures of data depth, local community depth captures local features of a distribution such as multimodality, and can adapt to modes with different scales. This gives our test greater power to detect scale changes in multimodal distributions, and our local depth is computationally cheaper in high dimensions than many classic depths like halfspace depth, regression depth, simplicial depth, and Mahalanobis depth. We test hypothesis by using local community depth in the Liu-Singh test, and we give conditions under which our method is consistent against fixed alternatives and has a limiting null distribution. In extensive simulations and an application to diagnosing dengue fever, we show that local community depth outperforms other depths when the underlying distribution is multimodal or asymmetric, and is comparable to other methods when the data is unimodal and symmetric.

Bayesian Analysis of GLMMs with Nonlocal Priors for Genome-Wide Association Studies
Marco Ferreira
Virginia Tech
marf@vt.edu
Coauthors: Shuangshuang Xu, Jacob Williams

We present a novel Bayesian method to find single nucleotide polymorphisms (SNPs) associated with particular phenotypes measured as discrete data from genome-wide association studies (GWAS). This is a regression problem with p two to three orders of magnitude larger than n, the subjects are correlated, and the SNPs regressors are highly correlated. To deal with these challenges, we propose nonlocal priors specifically tailored to GLMMs and develop related fast approximate computations for Bayesian model selection. To search through hundreds of thousands of possible SNPs, we use a two-step procedure: first, we screen for candidate SNPs; second, we perform model search that considers all screened candidate SNPs as possible regressors. A simulation study shows favorable performance of our Bayesian method when compared to other methods widely used in the GWAS literature. We illustrate our method with applications to the analysis of real GWAS datasets from plant science and human health.

Learning Communities in Data From Probabilistic Estimates of Similarity
John Foley
Metron, Inc.
foley@metsci.com
Coauthor: Kenneth S. Berenhaut, Wake Forest University

This talk describes an approach to infer community structure from probabilistic estimates of relative data similarity. The method generalizes the novel concept of cohesion introduced in Berenhaut, K. S., Moore, K. E., & Melvin, R. L. (2022) "A social perspective on perceived distances reveals deep community structure," Proceedings of the National Academy of Sciences, 119(4), e2003634119. The formulation we present builds on loc. cit. by distilling two key probabilistic concepts: conflict relevance, R, and support division, Q. We will recall the social motivation underlying the approach and describe new research directions enabled by our generalized formulation for applications with uncertainty in data similarity. Such applications include clustering with insights into internal cluster structure.

Deep Gaussian Process Surrogates for Computer Experiments
Robert Gramacy
Virginia Tech
rbg@vt.edu
Coauthors: Annie Sauer, Andy Cooper

Deep Gaussian processes (DGPs) upgrade ordinary GPs through functional composition, in which intermediate GP layers warp the original inputs, providing flexibility to model non-stationary dynamics. Recent applications in machine learning favor approximate, optimization-based inference for fast predictions, but applications to computer surrogate modeling – with an eye towards downstream tasks like calibration, Bayesian optimization, and input sensitivity analysis – demand broader uncertainty quantification (UQ). We prioritize UQ through full posterior integration in a Bayesian scheme, hinging on elliptical slice sampling the latent layers. We demonstrate how our DGP's non-stationary flexibility, combined with appropriate UQ, allows for active learning: a virtuous cycle of data acquisition and model updating that departs from traditional space-filling design and yields more accurate surrogates for fixed simulation effort. But not all simulation campaigns can be developed sequentially, and many existing computer experiments are simply too big for full DGP posterior integration because of cubic scaling bottlenecks. For this case we introduce the Vecchia approximation, popular for ordinary GPs in spatial data settings. We show that Vecchia-induced sparsity of Cholesky factors allows for linear computational scaling without compromising DGP accuracy or UQ. We

vet both active learning and Vecchia-approximated DGPs on numerous illustrative examples and a real simulation involving drag on satellites in low-Earth orbit. We showcase implementation in the deepgp package for R on CRAN.

The Interplay of Biomass Energy Consumption on Ecological Footprint: Using Parametric and Time-Varying Non-parametric Approaches

Shaymal Halder
University of Pennsylvania
haldersh@gvsu.edu
Coauthors: Shamal Chandra Karmaker, Kanchan Kumar Sen, Shahadat Hosan, Md. Matiar Rahman, Andrew Chapman, Bidyut Baran Saha, Grand Valley State University

The environmental implications of biomass energy use have become an emerging and debatable issue among policymakers. As biomass is one of the major traditional easy sources of energy, its impact on health and the economy is justified in several studies. However, evidence is scarce on the use of biomass energy in climate change mitigation. This study thus explores the repercussion of biomass energy consumption on the ecological footprint in OECD countries for the period 1990-2017. The existing literature is based on parametric approaches that can provide estimates for the average effect of biomass energy consumption on ecological footprint, however, they cannot demonstrate how the relationship has changed over time. Hence, the present study employed both parametric and nonparametric time-varying techniques to determine the impact of biomass energy use on the ecological footprint in studied countries over the time period. Results from both methods indicate that biomass energy consumption raises the ecological footprint of the OECD nations. Nonparametric findings suggest this relationship is time-varying. Policies are proposed to lessen the negative impact of biomass energy use on the environment based on these findings, exploring possible ways for other clean energy sources.

Statistical Consulting in a University Setting: Modern Challenges and Enduring Issues

Daniel Hall
University of Georgia
danhall@uga.edu
Coauthors: Marianne Huebner, Michigan State University; Krista Gile, University of Massachusetts, Amherst; Pat Gerard, Clemson University

Most research universities in the US have statistical consulting units of one form or another. These units are of great value to the academic units with which they are associated by enhancing graduate education, providing sources of financial support, and offering opportunities for collaborative research for students and faculty. They are also valuable to the broader university by bringing expertise in statistical and scientific methodology to research projects across a wide range of academic disciplines. Working as a statistical consultant and collaborator in the university environment is interesting, rewarding, and fun, despite and because of its many challenges. In this panel discussion, four directors of university statistical consulting units offer their experiences and insights regarding some of the perennial and modern challenges that arise in their consulting centers. Topics of discussion include (i) training students to be effective statistical consultants/collaborators; (ii) ethical challenges facing the statistical consultant/collaborator; (iii) how do changes in our discipline (e.g., a shift to a broader scope under the heading of "data science", an increasing emphasis on big data problems and methods in research and training) pose challenges for the university statistical consulting center, where small data problems remain dominant; and (iv) ensuring that a statistical consulting unit is valued by stakeholders and has sustainable funding for its long-term health.

Active Learning Meets Experimental Design Theory

Bill Heavlin
Google
bill.heavlin@gmail.com

Active learning (AL) algorithms seek to identify (large batches of) candidates to label, both to improve an underlying machine learning model and to harvest higher utility. Such algorithms typically key on candidates of high predicted utility while incorporating some ad hoc criterion for promoting diversity. In contrast, our algorithm ("EIO-AL") draws from the statistical theory of experimental design, resulting in a particularly elegant representation of diversity. With simple modifications, we sensitize this algorithm to favor higher utility candidates. On CIFAR-10, we achieve statistical efficiency equivalent to 4-5 times that of randomly selected candidates.

Mean Structure Learning with High-Dimensional Correlated Data

Emily Hector
North Carolina State University
ehector@ncsu.edu

Motivated by image-on-scalar regression with data aggregated across multiple sites, we consider a setting in which multiple independent studies each collect multiple dependent vector outcomes, with potential mean model parameter homogeneity between studies and outcome vectors. To determine the validity of jointly analyzing these data sources, we must learn which of these data sources share mean model parameters. We propose a new model fusion approach that delivers improved flexibility, statistical performance and computational speed over existing methods. Our proposed approach specifies a quadratic inference function within each data source and fuses mean model parameter vectors in their entirety based on a new formulation of a pairwise fusion penalty. We establish theoretical properties of our estimator and propose an asymptotically equivalent weighted oracle meta-estimator that is more computationally efficient. Simulations and application to the ABIDE neuroimaging consortium highlight the flexibility of the proposed approach.

Optimal Design for Ordinal Categorical Regression on Milk Fiber Strength

Yiren Hou
University of Georgia
yh82581@uga.edu

Milk and clothing industries produce substantial waste. However, these wasted or expired milk can be reused through its composition, a milk protein called casein. The milk fiber created from casein is a sustainable alternative to other textiles. This research aims to find the optimal process in making a strong milk fiber. Following environmental-friendly steps in creating the milk fiber, different additives are utilized in the dope that is used for fiber extrusion. These additives or predictors assigned with different weight concentrations are casein powder, cellulose nanofibril, beeswax, and corn protein zein. An orthogonal array-based Maximin design has been identified for this project. Experiments were conducted following this plan. Qualitative observations of the extruded fiber are recorded and are classified as ordinal categorical response with three levels for strength. Under ordinal logistic regression, beeswax and casein have greater effects on the strength, and cellulose nanofibril has a quadratic relationship with the strength of fiber.

Online Controlled Experiments: Top Challenges and Solutions

Jennie Huang
Duke University
yjh3@duke.edu
Coauthor: David Banks

Online controlled experiments are an effective way to evaluate the impact of changes made to software products, customer experiences, ad campaigns, and more [1]. In recent years, experimentation has become key to the success of companies, thousands of which have invested in tools such as Optimizely, Google Optimize, Mixpanel, VWO, AB Tasty, and Split.io [2]. Although the concept of an OCE is simple, there are quite a few challenges that come with running large-scale OCEs. This unique set of challenges offers opportunities for statisticians to get involved in this "modern playground." This talk will give an overview of online controlled experiments, present the top challenges of running OCEs, and discuss recent developments intended to address such challenges (including the multi-armed bandit, experimentation in the presence of networks etc.). We will also walk through some examples of OCEs done within large tech organizations such as Amazon, Uber, Google, and Twitter.

References

[1] Gupta, Somit, et al. "Top challenges from the first practical online controlled experiments summit." ACM SIGKDD Explorations Newsletter 21.1 (2019): 20-35.
[2] Stevens, Nathaniel. "Modern Design of Experiments in Computational Advertising." Statistical Methods for Computational Advertising, October 5th, 2021.

Quantile Foliation - Smoothing to Model Performance in Olympic Weightlifting Across the Life Span

Marianne Huebner
Michigan State University
huebner@msu.edu
Coauthor: Aris Perperoglou, AstraZeneca, UK

We developed an extension of quantile regression models and quantile sheets, called "quantile foliation" that is used to predict outcomes for one explanatory variable based on two covariates and varying quantiles. We study performances, as measured by the total weight lifted, from World Championships in Olympic weightlifting for athletes aged 13 to 90. With quantile foliation it is possible to examine age-associated patterns of performance increase for youth, and to study the decline after reaching the peak performance. This can be done for athletes with different body mass and performance levels as measured by quantiles. Novel contributions include a comparison of youth athletes' performances for different body mass, and age-associated performance decline for female Master athletes. Results of this work has led to a change in scoring rules in international weightlifting competitions.

Impact of Establishment of Liquid Medical Oxygen System Indifferent Health Facility to Provide Quality Care of Critical Patient During Covid-19 Crisis in Bangladesh

Mohammad Saiful Isla
DGHS, Ministry of Health and Family Welfare of Bangladesh
drsaifulislam@gmail.com
Coauthor: Habib

Introduction: Oxygen is an essential medicine that is used to treat hypoxaemia at all levels of the healthcare system, yet in the 21st century it is not available for many patients who require oxygen, especially in low-resource settings. Bangladesh experienced a surge in COVID-19 cases at an unprecedented rate since the last week of May 2021. As the number of cases continued to rise exponentially, scarce hospital resources ran thin, and critical care units were overburdened. During the second wave of COVID-19 pandemic,

Bangladesh experienced a surge in COVID-19 cases at an unprecedented rate since the last week of May 2021. COVID-19 case surge created extraordinary burden on the health system of the country. As oxygen therapy was used as the main treatment of moderate, severe, and critical COVID-19 patients, the use of medical oxygen raised exponentially during this surge. As a result, like other countries, Bangladesh also struggled to ensure the availability of medical oxygen during this surge. Being concerned, the Government of Bangladesh (GOB) stepped forward to improve production and distribution of medical oxygen for hospitals and thus improve the outcome of patients affected by COVID-19. MOHFW established Liquid Medical Oxygen System more than 100 health facilities to provide uninterrupted medical oxygen support for Covid and non Covid critical patients which was supported by World Bank with technical support of Unicef, Bangladesh.

Methodology : Data was collected on the basis of oxygen production, distribution, and consumption. Eight Review of documents on use of medical gas in health system, and demand and supply status, current practices on medical gas production, etc. Interview with key experts (including medicine specialists and critical care experts). meeting with key stakeholders and Visit to health facilities at national and peripheral levels.

Results: This study found that before COVID-19 pandemic, average oxygen consumption for the whole country was 35 tons per day (TPD). Maximum consumption was in Dhaka, followed by Chattogram and Rajshahi divisions. Dhaka division alone had around 47% consumption of the country during the pre-COVID-19 period. The average daily cost for oxygen is around 40 lac taka. During April 2020 - December 2021 (COVID-19 pandemic period), consumption of oxygen ranged from 22 to 220 ton per day. While the oxygen consumption was highest in Dhaka division, it also increased drastically in other divisions during the pandemic. During the peak in August 2021, Dhaka consumed 39%, Chattogram 22%, Rajshahi 23%, Khulna 7%, Barishal 5%, Rangpur 3%, and Sylhet 1% of the total oxygen consumption in the country. Peak consumption of oxygen during COVID-19 (205 TPD) was 5 times higher than normal (non-COVID) peak demand (39 TPD).Consumption of nitrogen, nitrous oxide and carbon di-oxide were very low and had no effect on consumption of these medical gases during the pandemic. The regular daily nitrogen consumption was average 0.88 ton during normal situation. Nitrous oxide and carbon di-oxide consumption was also very low. During COVID-19 period, daily average nitrogen consumption was found as 2.38 tons, nitrous oxide 0.32 ton and carbon di-oxide 0.068ton. The overall daily medical oxygen supply in the country, combined together with the output from oxygen producers and importers, was about 170 tons. The country's medical oxygen demand increased 10-fold after March 2021 which in no way could be met with the local production. The supplier company could every week import only 15 tons from India while their production was 33 tons daily against a capacity of 54 tons as technical problems deter the plant to operate in full capacity and daily medical oxygen supply rose to 50 tons from 30 tons in March 2021, adding that they were also importing the gas to meet the increasing demand.

Recommendation : Oxygen requirement of public hospitals is met through procuring oxygen from two local private manufacturers which are not sufficient .Being concerned, the Government of Bangladesh stepped forward to improve production and distribution of medical oxygen for hospitals and clinics and thus improve the outcome of patients affected by COVID-19 to meet the medical oxygen and other gas demand during case surge in future waves of COVID-19 pandemic or other epidemics, and fully meet the health system's need in non-pandemic times and support a strengthened health system for the whole population.

Quantifying the Response of Blainville's Beakedwhales to U.S. Naval Sonar Exercises in Hawaii

Eiren Jacobson
Centre for Research into Ecological and Environmental Modelling, University of St. Andrews, UK
ej45@st-andrews.ac.uk
Coauthors: E. Elizabeth Henderson, Naval Information Warfare Center Pacific, San Diego, California; David L. Miller, Centre for Research into Ecological and Environmental Modelling, School of Mathematics and Statistics, University of St Andrews, St Andrews, Scotland; Cornelia S. Oedekoven, Centre for Research into Ecological and Environmental Modelling, School of Mathematics and Statistics, University of St Andrews, St Andrews, Scotland; David J. Moretti, Naval Undersea Warfare Center, Newport, Rhode, Island; Len

Thomas, Centre for Research into Ecological and Environmental Modelling, School of Mathematics and Statistics, University of St Andrews, St Andrews, Scotland

Behavioral responses of beaked whales (family Ziphiidae) to naval use of mid-frequency active sonar (MFAS) have been quantified for some species and regions. We describe the effects of MFAS on the probability of detecting diving groups of Blainville's beaked whales on the U.S. Navy Pacific Missile Range Facility (PMRF) in Hawaii and compare our results to previously published results for the same species at the Atlantic Undersea Test and Evaluation Center (AUTEC) in the Bahamas. We use passive acoustic data collected at bottom-mounted hydrophones before and during six naval training exercises at PMRF along with modelled sonar received levels to describe the effect of training and MFAS on foraging groups of Blainville's beaked whales. We use a multistage generalized additive modeling approach to control for the underlying spatial distribution of vocalizations under baseline conditions. At an MFAS received level of 150dB re 1 μPa rms the probability of detecting groups of Blainville's beaked whales decreases by 77%, 95% CI [67%, 84%] compared to periods when general training activity was ongoing and by 87%, 95% CI [81%, 91%] compared to baseline conditions. Our results indicate a more pronounced response to naval training and MFAS than has been previously reported.

State-Space Models for Marine Mammal Populations

Eiren Jacobson
Centre for Research into Ecological and Environmental Modelling, University of St. Andrews, UK
eiren.jacobson@gmail.com

Statistical ecology is an interdisciplinary field working to better understand the natural world. One area of active research within statistical ecology is the estimation of abundance, trend, and population dynamics of marine mammal populations. Marine mammals are often studied to better understand past, present, and future anthropogenic impacts on populations and ecosystems. Marine mammals spend much of their time underwater, where they are difficult for humans to observe. However, there are many different survey techniques available for marine mammals populations, including visual, passive acoustic, and genetic studies. Each mode of data collection has its own advantages and disadvantages. Using empirical statistical models on these types of data can lead to incorrect inference on the population parameters of interest due to imperfect observation processes. State-space models, where the population process and observation process are considered separately, can be used to explicitly account for bias and uncertainty arising from the observation process. These models often rely on a mechanistic description of both the population and observation processes. Here, I discuss how different types of data can be integrated in state-space models for marine mammal populations. These models can improve inference on population parameters, and, where funding for field studies is limited, can guide data collection to focus on sensitive parameters. To illustrate different types of state-space models used for marine mammal populations, I present case studies from my own work to develop state-space models for grey seals, harbour seals, and beluga whales.

Robust Topological Classification with Applications in Firn Data Analysis

Matt Jester
UNC Greensboro
mwjester@uncg.edu
Coauthors: Jesse Domino, CUNY Graduate Center; Xiaoli Gao, UNCG; Sarah Day, College of William & Mary; Kaitlin Keegan, University of Nevada, Reno

In traditional models for image classification, it is necessary that a well categorized set of training data is achievable. However, image acquisition is often affected by different sources of noise. Particularly, in firn data analysis, micro-CT images may be mislabeled across different depths due to continuous percolation combined with temporal and spatial dependence. In this paper, we propose a method for robust image classification that utilizes the topological information of firn images from 10 different depths. Topological features are

important signatures in image classification, and they have innate robustness due to their rotation and scale invariance. When we systematically mislabeled images in the training set, existing models failed to make accurate predictions for the test data. However, when we utilized stochastic cross validation and weighting algorithms, our proposed algorithm was able to make accurate predictions, even when the mislabeling rate was as high as 40%. Our results show that the topological information is robust in the case of mislabeled images in the training set.

A Graphical Multi-Fidelity Gaussian Process Model, With Application to Emulation of Heavy-Ion Collisions

Irene Ji
Duke University
yi.ji@duke.edu
Coauthors: Simon Mak, Derek Soeder, J-F Paquet, Steffen A. Bass

With advances in scientific computing and mathematical modeling, complex scientific phenomena such as galaxy formations and rocket propulsion can now be reliably simulated. Such simulations can however be very time-intensive, requiring millions of CPU hours to perform. One solution is multi-fidelity emulation, which uses data of different accuracies (or fidelities) to train an efficient predictive model which emulates the expensive simulator. For complex scientific problems and with careful elicitation from scientists, such multi-fidelity data may often be linked by a directed acyclic graph (DAG) representing its scientific model dependencies. We thus propose a new Graphical Multi-fidelity Gaussian Process (GMGP) model, which embeds this DAG structure (capturing scientific dependencies) within a Gaussian process framework. We show that the GMGP admits a scalable algorithm for recursive computation of the posterior mean and variance along sub-graphs. We also present an experimental design methodology over the DAG given an experimental budget, and propose a nonlinear extension of the GMGP model via deep Gaussian processes. The advantages of the GMGP model are then demonstrated via a suite of numerical experiments and an application to emulation of heavy-ion collisions, which can be used to study the conditions of matter in the Universe shortly after the Big Bang.

Testing for Irrelevance in Partially Parametric Models With Parametric Nulls

Jiancheng Jiang
University of North Carolina at Charlotte
jjiang1@uncc.edu
Coauthors: Daniel J. Henderson, Department of Economics, Finance and Legal Studies, University of Alabama

In this paper we consider tests of relevance for a partially parametric model. Specifically, we test that the entire nonparametric function is irrelevant in the prediction of the outcome variable. This test results in a parametric model which can be estimated via non-linear least-squares. One useful application of the aforementioned test is to test for the relevance of a control function in simultaneous equation models.

A Step-Wise Multiple Testing With Linear Regression Models for the Study of Resting Energy Expenditure

Zhezhen Jin
Columbia University
zj7@cumc.columbia.edu
Coauthors: Junyi Zhang, Zimian Wang, Zhezhen Jin, Zhiliang Ying

A new multiple hypothesis testing approach to evaluate organ/tissue-specific resting metabolic rates will be presented. The approach is based on generalized marginal regression estimates for a subset of coefficients along with a stepwise multiple testing procedure with a minimization-maximization of the normalized estimates (maximization over all its components and minimization over all possible choices of the subset). The approach offers a valid way to address challenges in multiple hypothesis testing on regression coefficients in linear regression analysis especially when covariates are highly correlated.Energy is of great importance to support normal metabolic functions, growth and repair of tissues and physical activity.

To Screen or Not to Screen? Using Data From Randomized Screening Trials to Quantify Risks Benefits of Cancer Screening

Karen Kafadar
University of Virginia
kk3ab@virginia.edu
Coauthors: Philip C. Prorok, National Cancer Institute

Cancer screening is assumed to be beneficial, in terms of reduced mortality and extended survival. Survival is often measured as the time between clinical detection of disease and endpoint (cure or death). When the disease is screen-detected, survival has two additional components: lead time (time by which the screening test advances the time of clinical diagnosis) and benefit time (extended survival time if the screen detection is beneficial). All three components are affected by two effects: length biased sampling (slow-growing cases are more likely to be screen-detected than fast-growing ones) and overdiagnosis (cases that are screen-detected but would never have surfaced clinically in the absence of screening). We quantify both effects in this talk and illustrate their non-trivial impacts on the results from actual randomized cancer screening trials.

Dimension Reduction for Gaussian Process Models via Convex Combination of Kernels

Lulu Kang
Illinois Institute of Technology
lkang2@iit.edu
Coauthor: Minshen Xu

Some engineering and scientific computer models that have high dimensional input space are actually only affected by a few essential input variables. If these active variables are identified, it would reduce the computation in the estimation of the Gaussian process (GP) model and help researchers understand the system modeled by the computer simulation. More importantly, reducing the input dimensions would also increase the prediction accuracy, as it alleviates the "curse of dimensionality" problem. In this talk, we propose a new approach to reduce the input dimension of the Gaussian process model. Specifically, the proposed optimization method iterates between adding kernels of lower dimension from a large candidate set of kernels to identify a convex combination of kernels and updating the weights, which can be considered as a Fedorov-Wynn type of algorithm. The combination of kernels is the correlation function for the GP model. To make sure a sparse subset is selected, we add the heredity principle while selecting the active input dimensions. Several numerical examples are shown to show the advantages of the method. The proposed method has many connections with the existing methods including active subspace, additive GP and composite GP models in the Uncertainty Quantification literature.

An Analysis of Demographic Trends Using Topological Data Analysis
Jakini Kauba
Clemson University
jkauba@g.clemson.edu
Coauthor: Thomas Weighill, UNCG

In recent years, Topological Data Analysis (TDA) has been used to analyze complex data and provide insights that other research techniques cannot. TDA is a newer form of data analysis which analyzes trends of data from a topological perspective by way of the main visualization tool of persistence diagrams. TDA has been used to measure breast cancer transcriptional DNA, voting patterns in precincts, gerrymandering, and even texture representation. In this paper, we apply TDA to geospatial data from the census to more accurately describe racial segregation among the Black and Hispanic demographics across one hundred cities in America. Our goal was to complete city to city comparisons in 2010 and 2020 as well as compare city similarities over the course of ten years for each race and note the respective trends. We were able to find seven clusters of cities in the black population that shared common characteristics and five for the Hispanic population. After doing a comparison of cities across the span of a decade, we also found commonalities of each racial demographic. In summary, this project represents a first step in uncovering trends in demographic data using TDA. We hope to continue exploring this data set in an effort to expand our understanding of various demographic patterns in America.

Statistical Disclosure Risk with Differential Privacy, with Application to the 2020 Decennial Census
Zekican Kazan
Duke University
zekican.kazan@duke.edu
Coauthor: Jerome P. Reiter

We propose Bayesian methods to assess the statistical disclosure risk of data released under differential privacy, focusing on settings with a strong hierarchical structure. The risk assessment is performed by hypothesizing Bayesian intruders with various amounts of prior information and examining the distance between their posteriors and priors. We discuss applications of these risk assessment methods to differentially private data releases from the 2020 decennial census and perform simulation studies using public individual-level data from the 1940 decennial census. Among these studies, we examine how the data holder's choice of privacy parameter affects the disclosure risk and quantify the increase in risk when a hypothetical intruder incorporates substantial amounts of hierarchical information.

Actively Learning About Active Learning
Timothy Keaton
Purdue University
keatont@purdue.edu
Coauthor: Arman Sabbaghi

The field of active learning resides at the intersection of experimental design and machine learning, and it has strong connections to traditional concepts in the design of experiments such as response surface methodology and optimum experimental design. The field has seen a tremendous surge in popularity and usage in modern applications. In this presentation we will provide an overview of active learning and provide examples of active learning methodologies and algorithms in practice. This presentation will provide the necessary background for understanding the other talks in this session, as well as a forum to promote a broader discussion of active learning at AISC 2022. We will encourage audience questions and discussion about this exciting field during our presentation.

Mean Estimation Using RRT Models Under Two-Phase Simple & Stratified Random Sampling Designs

Sadia Khalil
UNC Greensboro & Lahore College for Women University
s_khali2@uncg.edu
Coauthor: Hafiza Fakhar ul Nisa, Lahore College for Women University

We have suggested the Generalized RRT mean estimators employing full and optional RRT models to examine the effects of measurement error and non-response error on estimation of the mean of a sensitive study variable under two-phase simple and stratified random sampling. These estimators are derived under four different scenarios of both errors. The Performance of proposed classes of estimators is evaluated using a unified measure of efficiency and privacy protection. Through extensive simulations, the validity of theoretical results is examined.

On Estimation and Monitoring of Population Mean Using Generalized Neutrosophic Ratio-Type Exponential Estimator Under Neutrosophic Exponentially Weighted Moving Average Scheme

Hina Khan
Government College University Lahore, Pakistan
hinakhan@gcu.edu.pk
Coauthors: Zaigham Tahir and Yasar Mahmood, Government College University Lahore, Pakistan

Classical studies under crisp and definite information employing the relationship of the research variable with auxiliary information to estimate the finite population mean may be seen in the past. Neutrosophic statistics (generalized classical statistics) has grown in popularity in recent years. It deals with indeterminacy, ambiguity, and uncertainty in data. The neutrosophic observation attains the form $O_N = O_L + O_U I_N$, where $I_N \in [I_L, I_U], O_N \in [O_L, O_U]$. Because the significance of statistical quality control is obvious. This study is the first work in neutrosophic statistics to employ a novel generalized neutrosophic ratio-type exponential estimator (NRTEE) to generate a control chart for the location parameter (mean) under interval neutrosophic data (IND). This estimator can also estimate the mean of a finite population using auxiliary data. We have examined the Neutrosophic exponentially weighted moving average (NEWMA) control chart (CC) utilizing NRTEE under IND which may be called as NEWMA-NRE CC for the various in-control average run-length (ARL) properties with various combinations of smoothing constant λ_N. We also have examined how changes in the values of λ_N effects ARL's properties. To get convincing evidence in favor of the NEWMA-NRE CC, we have used the neutrosophic Monte Carlo simulations. Under the classic ratio estimator, our NEWMA-NRTEE CC outperforms the existing EWMA CC. Our chart fared effectively in recognizing an out-of-control operation for minor shifts.

Application of Modified Systematic Sampling in Auto-correlated Populations

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

In this study, the selection procedure of modified systematic sampling is improved with an alternative approach which is relatively simple and easy to understand. This scheme is more generalized and not only restricted to the case where population size is multiple of sample size but also applicable for all possible combination of population size and sample size. Numerical efficiency comparison of modified systematic sampling with the well-known sampling schemes has been carried out for auto-correlated model using linear, exponential and hyperbolic correlogram.

Unbiased Estimation of Variance of Sample Mean in Systematic Sampling
Zaheen Khan
Federal Urdu University of Arts, Science and Technology, Islamabad
zaheen.khan@fuuast.edu.pk

The problem of unbiased estimation of variance of sample mean/total in systematic sampling has always been widely discussed by the researchers and practitioners. The main difficulty in unbiased estimation of variance of sample mean/total arises due to the reason that second-order inclusion probabilities are not positive in usual systematic sampling. In this paper, we propose an alternative systematic sampling scheme which ensures an unbiased estimation of the variance of sample mean or total. Finally, a detailed empirical efficiency comparison has also been carried out in this paper.

Computing Generalized Rank Invariant via Zigzag Persistence
Woojin Kim
Duke University
woojin@math.duke.edu
Coauthors: Tamal Dey, Purdue University; Facundo Mémoli, The Ohio State University

The notion of generalized rank invariant in the context of multiparameter persistence has become an important ingredient for defining interesting homological structures such as generalized persistence diagrams. Naturally, computing these rank invariants efficiently is a prelude to computing any of these derived structures efficiently. We show that the generalized rank over an interval I of a 2-paramater persistence module M is equal to the generalized rank of the zigzag module that is induced on a certain path in I tracing mostly its boundary. Hence, we can compute the generalized rank over I by computing the barcode of the zigzag module obtained by restricting the bifiltration inducing M to that path. Among others, we apply this result to obtain an improved algorithm for the following problem. Given a bifiltration inducing a module M, determine whether M is interval decomposable and, if so, compute all intervals supporting its summands.

Latent Gaussian Dynamic Factor Modeling and Forecasting for Multivariate Count Time Series
Younghoon Kim
UNC Chapel Hill
yhkim0225@unc.edu
Coauthors: Zachary F. Fisher, The Pennsylvania State University; Vladas Pipiras, University of North Carolina at Chapel Hill

This work studies estimation and forecasting of multivariate count time series models driven through a copula type transformation by a Gaussian dynamic factor model. The estimation is based on second-order properties of the count and underlying Gaussian models, and applies to the case where the model dimension is larger than the sample size. The forecasting is carried out through a particle-based sequential Monte Carlo, leveraging Kalman filtering techniques. A simulation study and an application are also considered.

Development of the Subject Statistics: A Historical Perspective

Sheela Kumari
Department of History, BPSIHL, BPSMV, Khanpur, Sonepat (India)
drsheelabpsmv@gmail.com

The purpose of this article is to present a brief historical perspective on the growth path of the subject Statistics during 17th to 20th century (1650-1950). Some important key development of the subject statistics have been highlighted to aware the researchers of different streams including social sciences, management and engineering. The information about the world renowned societies and research journals has been provided to emphasise on the historical growth path of the subject statistics. In particular the contribution of the father of Modern Statistics in India (P.C. Mahalanobis) has also been addressed to reflect the origin and development of the subject Statistics in India.

Keywords: Historical Perspective, Development of Statistics, Renowned Societies, Research Journals and Contribution of P.C. Mahalanobis

A Scalable Method for Fitting Sparse Markov Models

Soumendra Lahiri
North Carolina State University
snlahiri@ncsu.edu

We consider sparse Markov models (SMMs) where the transition probabilities of an m-th order Markov chain are identical over sets in a partition of the m-th order past histories. The number of potential partitions grows at a super-exponential rate with the order m (given by the Bell number); the problem we consider here is to select the true model from this large collection with high probability. We propose a simple and scalable method for identifying the partitions and investigate its theoretical properties. This is a joint work with Donald Martin and Tuhin Majumder.

Cross-Domain Recommender Systems

Patrick LeBlanc
Duke University
pml26@duke.edu
Coauthors: Timothy C. Au; David Banks, Duke University; Linhui Fu, Duke University; Mingyan Li, UNC Greensboro; Zhengyu Tang, Duke University; Qiuyi Wu, University of Rochester

Recommender systems are the engine of on-line advertising. Not only do they suggest movies, music, or romantic partners, but they also are used to select which advertisements to show to users. Traditionally, recommender systems have restricted themselves to only one such domain: movie recommenders only consider user preferences on movies. Cross-domain recommender systems, in contrast, attempt to use what has been learned about a user in one context to inform recommendations in a different context - leveraging, for example, book preferences to learn movie preferences. We cover an existing method extending probabilistic matrix factorization (PMF) to the multi-domain context and propose a multi-domain extension of the Bayesian PMF model.

Generalizing the German Tank Problem

Anthony Lee
Milton Academy
anthonyjlee0101@gmail.com
Coauthor: Steven J Miller, Wiliams College

The German Tank Problem dates back to World War II when the Allies used a statistical approach to estimate the number of enemy tanks produced or on the field from observed serial numbers after battles. In the original problem, assuming that the tanks are labeled consecutively starting from 1, if we observe k tanks from a total of N tanks with the maximum observed tank being m, then the best estimate for N is $m(1 + 1/k) - 1$. First, we looked at the discrete and continuous one dimensional case. We attempted to improve the original formula by using different estimators such as the second largest and L^{th} largest tank, and applied motivation from portfolio theory by seeing if a weighted average of different estimators would produce less variance; however, the original formula, using the largest tank proved to be the best. The continuous case was similar, as we proved that using just the largest tank in the continuous setting gives the best estimate for N. Then, we attempted to generalize the problem into two dimensions, where we pick pairs instead of points. We looked at the discrete and continuous square and circle variants. There were more complications in the two dimensional problem than in the original problem, as we dealt with problems in geometry and number theory such as dealing with curvature issues in the circle, and the problem that not every number is representable as a sum of two squares. In some cases, we concentrated on the large N limit (with fixed k) by deriving approximate formulas by keeping only the main term of the computation and then inverting, due to the complexity of the exact formulas and the lack of significant gain they provided over the approximate ones. For the discrete and continuous square, we tested various statistics, but found that the largest observed component of our pairs is the best statistic to look at; the scaling factor for both cases is $(2k + 1)/2k$. For the circle we used motivation from the equation of a circle; for the continuous case, we looked at $\sqrt{X^2 + Y^2}$ and for the discrete case, we looked at $X^2 + Y^2$ and took a square root at the end to estimate for r. The discrete case was especially involved because we had to use approximation formulas that gave us the number of lattice points inside the circle. Interestingly, the scaling factors were different for the cases. Lastly, we generalized the problem into L dimensions squares and circles. The discrete and continuous square proved similar to the two dimensional square problem. However, for the L^{th} dimensional circle, we had to use formulas for the volume of the L-ball, and had to approximate the number of lattice points inside it. The formulas for the discrete circle were particularly interesting, as there was no L dependence in the formula. After the L-dimensional case, we concluded the paper with an appendix with detailed information on portfolio theory, omitted calculations, and Mathematica code to simulate the German Tank problem.

A Scalable Partitioned Approach to Model Massive Nonstationary Non-Gaussian Spatial Datasets

Ben Seiyon Lee
George Mason University
slee287@gmu.edu
Coauthor: Jaewoo Park, Department of Statistics and Data Science, Yonsei University

Nonstationary non-Gaussian spatial data are common in many disciplines, including climate science, ecology, epidemiology, and social sciences. Examples include count data on disease incidence and binary satellite data on cloud mask (cloud/no-cloud). Modeling such datasets as stationary spatial processes can be unrealistic since they are collected over large heterogeneous domains (i.e., spatial behavior differs across subregions). Although several approaches have been developed for nonstationary spatial models, these have focused primarily on Gaussian responses. In addition, fitting nonstationary models for large non-Gaussian datasets is computationally prohibitive. To address these challenges, we propose a scalable algorithm for modeling such data by leveraging parallel computing in modern high-performance computing systems. We partition the spatial domain into disjoint subregions and fit locally nonstationary models using a carefully curated set of spatial basis functions. Then, we combine the local processes using a novel neighbor-based weighting scheme. Our approach scales well to massive datasets (e.g., 2.7 million samples) and can be

implemented in nimble, a popular software environment for Bayesian hierarchical modeling. We demonstrate our method to simulated examples and two massive real-world datasets acquired through remote sensing.

A Family of Orthogonal Main Effects Screening Designs for Mixed Level Factors

Ryan Lekivetz
SAS
ryan.lekivetz@jmp.com
Coauthors: Bradley Jones, JMP; Christopher Nachtsheim, Carlson School of Management

There is scant literature on screening when some factors are at three levels and others are at two levels. Two well-known and well-worn examples are Taguchi's L18 and L36 designs. However, these designs are limited in two ways. First, they only allow for either 18 or 36 runs, which is restrictive. Second, they provide no protection against bias of the main effects due to active two-factor interactions (2FIs). In this talk, we introduce a family of orthogonal, mixed-level screening designs in multiples of eight runs. Our 16-run design can accommodate up to four continuous three-level factors and up to eight two-level factors. The two-level factors can be either continuous or categorical. All of our designs supply substantial bias protection of the main effects estimates due to active 2FIs.

Probabilistic Contrastive Principal Component Analysis

Didong Li
UNC Chapel Hill
didongli@unc.edu

Dimension reduction is useful for exploratory data analysis. In many applications, it is of interest to discover a variation that is enriched in a "foreground" dataset relative to a "background" dataset. Recently, contrastive principal component analysis (CPCA) was proposed for this setting. However, the lack of a formal probabilistic model makes it difficult to reason about CPCA and tune its hyperparameter. We propose probabilistic contrastive principal component analysis (PCPCA), a model-based alternative to CPCA. We discuss how to set the hyperparameter in theory and in practice, and we show several of PCPCA's advantages over CPCA, including greater interpretability, uncertainty quantification and principled inference, robustness to noise and missing data, and the ability to generate data from the model. We demonstrate PCPCA's performance through a series of simulations and case-control experiments with datasets of gene expression, protein expression, and images.

Variational Manifold-Embedded Gaussian Process Modeling, With Applications to Aircraft Engine Design

Kevin Li
Duke University
kevin.li1324@gmail.com
Coauthors: Simon Mak, Duke University - Department of Statistical Science; Matthew Plumlee, Northwestern University - Department of Industrial Engineering and Management Sciences; Suo Yang, University of Minnesota - Department of Mechanical Engineering

Gaussian processes (GPs) provide a flexible non-parametric Bayesian framework for probabilistic predictive modeling. It is, however, well-known that such models may suffer from a "curse-of-dimensionality", in that its performance can deteriorate greatly as input dimensionality increases. One saving grace is that, for many problems, the underlying response surface is known to be active only on a low-dimensional manifold embedding. In physical science application, such embeddings often capture a sparse number of dominant physics which dictate the high-dimensional system, and the integration and uncertainty quantification of

such embeddings is thus crucial for predictive scientific computing. We propose a new Variational Manifold-Embedded GP (VME-GP) model, which leverages a variational inference approach for probabilistic learning of such embeddings within a GP framework. In particular, the VME-GP makes use of carefully-specified shrinkage priors and optimization on the embedded Steifel manifold, which allows for an efficient and probabilistic integration of this low-dimensional structure within the GP model. We demonstrate the effectiveness of the VME-GP over existing methods in a suite of high-dimensional numerical experiments and for a turbomachinery case study involving the design of an aircraft engine fan blade.

Probabilistic Factorization Matrix

Mingyan Li
UNC Greensboro
m_li6@uncg.edu

Recommender system has been a hot topic in both industry and academia. A recommender system tries to predict the user preferences through utility functions based on the user, item, and preference, information, depending on the particular method being used. If you have bought anything from Amazon.com, you have experienced recommender system in the "Customer who bought this item also bought..." section; If you have a Netflix subscription, you have experienced recommender system when you see "Top Picks for XXX" or "Because you watched XXX". A lot of recommender system methods are inspired by the Netflix Prize competition, including the topic of the talk, Probabilistic Matrix Factoriztion (PMF). A PMF is a collaborative filtering algorithm, one of the main methods of recommender system, which uses only the preference information to predict user preference over some items. The talk is going to introduce the original PMF method by Mnih and Salakhutdinov in 2007, and some of its derivatives.

Decomposition of Variation of Mixed Variables by a Latent Mixed Gaussian Copula Model

Quefeng Li
UNC Chapel Hill
quefeng@email.unc.edu
Coauthors: Yutong Liu, Toni Darville, Xiaojing Zheng

Many biomedical studies collect data of mixed types of variables from multiple groups of subjects. Some of these studies aim to find the group-specific and the common variation among all these variables. Even though similar problems have been studied by some previous works, their methods mainly rely on the Pearson correlation, which cannot handle mixed data. To address this issue, we propose a latent mixed Gaussian copula (LMGC) model that can quantify the correlations among binary, ordinal, continuous, and truncated variables in a unified framework. We also provide a tool to decompose the variation into the group-specific and the common variation over multiple groups via solving a regularized M-estimation problem. We conduct extensive simulation studies to show the advantage of our proposed method over the Pearson correlation-based methods. We also demonstrate that by jointly solving the M-estimation problem over multiple groups, our method is better than decomposing the variation group by group. We also apply our method to a Chlamydia trachomatis genital tract infection study to demonstrate how it can be used to discover informative biomarkers that differentiate patients.

Individualized Treatment Regimes Incorporating Imaging Features
Xinyi Li
Clemson University
lixinyi@clemson.edu
Coauthor: Michael Kosorok, University of North Carolina at Chapel Hill

Precision medicine seeks to discover an optimal personalized treatment plan and thereby provide informed and principled decision support, based on the characteristics of individual patients. With recent advancements in medical imaging, it is crucial to incorporate patient-specific imaging features in the study of individualized treatment regimes. We propose a novel, data-driven method to construct interpretable image features which can be incorporated, along with other features, to guide optimal treatment regimes. The proposed method treats imaging information as a realization of a stochastic process, and employs smoothing techniques in estimation. We show that the proposed estimators are consistent under mild conditions. The proposed method is applied to a dataset provided by the Alzheimer's Disease Neuroimaging Initiative.

Trusted Aggregation (TAG): Model Filtering Backdoor Defense In Federated Learning
Yao Li
UNC Chapel Hill
yaoli@email.unc.edu
Coauthors: Joseph Lavond, UNC Chapel Hill; Minhao Cheng, Hong Kong University of Science and Technology

Federated Learning is a framework for training machine learning models from multiple local data sets without access to the data. A shared model is jointly learned through an interactive process between server and clients that combines locally learned model gradients or weights. However, the lack of data transparency naturally raises concerns about model security. Recently, several state-of-the-art backdoor attacks have been proposed, which achieve high attack success rates while simultaneously being difficult to detect, leading to compromised federated learning models. In this paper, motivated by differences in the output layer distribution between models trained with and without the presence of backdoor attacks, we propose a defense method that can prevent backdoor attacks from influencing the model while maintaining the accuracy of the original classification task.

Bayesian Gaussian Copula Graphical Model for Ordinal Data
Xiaoyan (Iris) Lin
University of South Carolina
lin@stat.sc.edu
Coauthors: Yang He, University of South Carolina; Qun Zhao, Nanjing University of Information Science and Technology

Collecting survey responses is an effective way to gather data and to help to make decisions. In a survey, there are often an abundant number of survey questions and many questions have ordinal responses. Knowing the conditional dependence among questions can help to condense the survey by removing duplicated or unnecessary questions. However, it is challenging to access conditional dependence among a large set of discrete random variables. In this project, we apply Bayesian Gaussian copula graphical model to estimate the conditional dependence for ordinal variables. Following the idea of graphical Lasso prior, spike-and-slab Lasso prior is proposed for the regularization purpose. A block Gibbs sampling scheme is then developed for the posterior computation. Simulation studies compare the performance of using different priors: the graphical Lasso prior, adaptive graphical Lasso prior, and the spike-and-slab Lasso prior, and show a good estimation performance when using the adaptive graphical Lasso prior and the spike-slab Lasso prior. We then utilize the proposed methods to analyze two questions (one question has 54 sub-questions; the other has 10 sub-questions) in a survey, which is about the physiological and psychological health of current Chinese

undergraduate students. The graphical structure of the two questions is determined based on the estimated partial correlations among the latent Gaussian variables. The design of these two questions is then judged based on the estimated conditional dependence.

Fusion Learning: Combine Inferences From Diverse Data Sources

Regina Liu
Rutgers University
rliu@stat.rutgers.edu

Advanced data collection technology nowadays has often made inferences from diverse data sources easily accessible. Fusion learning refers to combining inferences from multiple sources or studies to make more effective inference than from any individual source or study alone. We focus on the tasks: 1) Whether/When to combine inferences? 2) How to combine inferences efficiently if we need to?

We present a general framework for nonparametric and efficient fusion learning for inference on multi-parameters, which may be correlated. The main tool underlying this framework is the new notion of depth confidence distribution (depth-CD), which is developed by combining data depth, bootstrap and confidence distributions. We show that a depth-CD is an omnibus form of confidence regions, whose contours of level sets shrink toward the true parameter value, and thus an all-encompassing inferential tool. The approach is shown to be efficient, general and robust. It readily applies to heterogeneous studies with a broad range of complex and irregular settings. This property also enables the approach to utilize indirect evidence from incomplete studies to gain efficiency for the overall inference. The approach will be shown with simulation studies and aircraft landing performance data.

This is joint work with Dungan Liu of University of Cincinnati and Minge Xie of Rutgers University.

Fitting Sparse Markov Models Through Regularization

Tuhin Majumder
Duke University
tmajumd@ncsu.edu

The major problem of fitting a higher order Markov model is the exponentially growing number of parameters. The most popular approach is to use a Variable Length Markov Chain (VLMC), which determines relevant contexts (recent pasts) of variable orders and form a context tree. A more general approach is called Sparse Markov Model (SMM), where all possible histories of order m form a partition so that the transition probability vectors are identical for the histories belonging to a particular group. We develop an elegant method of fitting SMM using convex clustering, which involves regularization. The regularization parameter is selected using BIC criterion. Theoretical results demonstrate the model selection consistency of our method for large sample size. Extensive simulation studies under different set-up have been presented to measure the performance of our method. We apply this method to classify genome sequences, obtained from individuals affected by different viruses.

Design and Analysis of Multi-Stage Multi-Fidelity Computer Experiments, With Application to Emulation of Heavy-Ion Collisions

Simon Mak
Duke University
sm769@duke.edu
Coauthors: Ruda Zhang, University of Houston; David Dunson, Duke University

In an era where scientific experimentation is costly, multi-fidelity emulation provides a powerful tool for predictive scientific computing. While there has been work on multi-fidelity modeling, existing models do not incorporate an important multi-stage property of simula- tors, where multiple fidelity parameters control for accuracy at different stages. We thus pro- pose a new Multi-stage Multi-fidelity Gaussian Process (M2GP) model, which embeds this multi-stage structure within a novel non-stationary covariance function, thus capturing prior knowledge on the numerical convergence of multi-stage simulators. Using the M2GP model, we then present a novel non-uniform COst-cOnstrained muLti-stage (COOL) designs, which aim to maximize information on the emulator model given experimental cost constraints. We investigate the predictive performance of the M2GP model and its associated COOL designs in a suite of numerical experiments and two applications, the first for emulation of cantilever beam deflection and the second for emulating the evolution of the quark-gluon plasma, which was theorized to have filled the Universe shortly after the Big Bang.

On Use of Regression Approach for Reliability Variation of a Parallel-Series System

Suresh Chander Malik
M D University Rohtak, India
scmalik@rediffmail.com

The role of structural design of components has been considered very significant for improving reliability of operational systems. Over the years a lot of research work has been made available by the engineers and scientist in the area of reliability theory and practice. As a result of which, several reliability improvement techniques such as configuration of the components in series, parallel, series-parallel, parallel-series and the provision of components in standby mode have been evolved. The researchers have claimed that a parallel system has more reliability than the series system and reliability of the systems can be improved much more by providing component wise redundancy. The reliability measures of systems with series-parallel and parallel-series configurations of the components have also been obtained by the researchers. On the other hand any change in structural design of a system invites lot of manipulations in terms of space and costs. But sometimes we need to reduce or increase the size of the components in a particular structure and in that situation it becomes very difficult to determine the reliability of the transformed systems that involves lot of computational efforts and costs. Therefore, in the present talk a method is suggested which can help the system developers to know the reliability of the transformed parallel-series system in terms of reliability of the original system and variation in reliability due to change in the structure. Here, the reliability of the transformed system obtained from the parallel-series system is discussed. And, the effect of failure rate of the components, number of parallel paths (with addition or removal of components) and number of components added (or removed) to (or from) the original system on variation in reliability has been examined using regression approach. The goodness of fit of the model has been checked on the basis of values of R^2 and adjusted R^2. The results are shown graphically and practical application of the work has been highlighted.

Modeling and Active Learning for Experiments with Quantitative-Sequence Factors
Abhyuday Mandal
University of Georgia
amandal@stat.uga.edu
Coauthors: Qian Xiao, Yaping Wang, Xinwei Deng

A new type of experiment that aims to determine the optimal quantities of a sequence of factors is eliciting considerable attention in medical science, bioengineering, and many other disciplines. Such studies require the simultaneous optimization of both quantities and the sequence orders of several components which are called quantitative-sequence (QS) factors. Given the large and semi-discrete solution spaces in such experiments, efficiently identifying optimal or near-optimal solutions by using a small number of experimental trials is a nontrivial task. To address this challenge, we propose a novel active learning approach, called QS-learning, to enable effective modeling and efficient optimization for experiments with QS factors. QS-learning consists of three parts: a novel mapping-based additive Gaussian process (MaGP) model, an efficient global optimization scheme (QS-EGO), and a new class of optimal designs (QS-design). The theoretical properties of the proposed method are investigated, and optimization techniques using analytical gradients are developed. The performance of the proposed method is demonstrated via a real drug experiment on lymphoma treatment and several simulation studies.

Smoothing Kernels for Categorical and Mixed-Scale Data
Marianthi Markatou
University of Buffalo
markatou@buffalo.edu

Kernels are essential elements in the construction of learning systems and have received considerable attention in machine learning. In statistics, kernels are used as tools for achieving specific data analytic goals, such as density estimation.

We discuss the construction and properties of a special class of kernels, the class of diffusion kernels. We first offer a statistical definition of this class, and present an important sub-class, the set of canonical diffusion kernels. We next present an algorithm to construct kernels for categorical scale data, either nominal or ordinal, and extend this construction to obtain kernels appropriate for use with mixed-scale data, that is both categorical and interval scale data. Our algorithm uses ideas that relate to the theory of continuous time Markov processes and the theory of Toeplitz matrices. We illustrate the construction of these kernels in high-dimensional density estimation. Time permitting, we will indicate the construction of tests statistics, akin to chi-squared tests of independence.

Random Persistence Diagram Generator
Vasileios Maroulas
University of Tennessee Knoxville
vmaroula@utk.edu

Topological data analysis (TDA) studies the shape patterns of data. Persistent homology (PH) is a widely used method in TDA that summarizes homological features of data at multiple scales and stores them in persistence diagrams (PDs). In this talk we will discuss a random persistence diagram generation (RPDG) method that produces a sequence of random PDs based on the PDs of the original data. RPDG is underpinned by (i) a model based on pairwise interacting point processes for inference of persistence diagrams, and (ii) by a reversible jump Markov chain Monte Carlo (RJ-MCMC) algorithm for generating samples of PDs. An example on a materials science problem will demonstrate the applicability of the RPDG method.

Inference for Hidden Sparse Markov Models

Donald E.K. Martin
North Carolina State University
demarti4@ncsu.edu
Coauthor: Iris Bennett, Corteva Agriscience

Whereas the number of parameters in a general higher-order Markov model is exponential in the order of dependence and the model has limited flexibility, sparse Markov models help with these problems. A sparse Markov model is a higher-order Markov model for which conditioning histories are grouped into classes such that the conditional probability distribution given any history of the class is the same. We introduce a model where variables following a sparse Markov structure are latent, and all inference over the latent states is conditioned on observed data. Then several tasks are considered in this sparse Markov setting: determining an appropriate model and parameter estimation, methodology for efficient computation of conditional distributions of pattern statistics over the hidden states, determining the likelihood of the observations, and obtaining the most likely hidden state at each time point and the most likely hidden state sequence, given the observations. An application is given to modeling the fluctuations in price of the S&P 500.

What's the Big Deal With Data Ethics, and Why Should I Care?

Wendy Martinez
U.S. Census Bureau
wendy.l.martinez@census.gov

You've probably heard the saying "Data is the new oil" (Clive Humby, 2006). Regardless of whether this is a good metaphor for how we view data today, we can safely assert that data is at the foundation of what we do as statisticians and data scientists. After all, our analyses and models all start with data and a problem to be solved. We need to be aware of ethical considerations with respect to the data throughout the data lifecycle. In this presentation, I will describe data ethics and provide motivating examples of where the data used had a negative ethical impact on the results. As a group, we will explore some case studies to further illustrate the concepts.

Cancer Data Science: Drug Testing in Cancer Research Using Auxiliary Information

Sunil Mathur
Houston Methodist, Weill Cornell Medical College
skmathur111@yahoo.com

The majority of metaplastic breast cancers (MpBC) Human epidermal growth factor receptor 2 (HER2)-negative Hormone receptor-negative, and a minority have low levels of staining for hormone receptors. The current standard of care for first-line HER2-negative metastatic or locally advanced MpBC follows the guidelines for the treatment of Triple Negative Breast Cancer (TNBC) chemotherapy with immune checkpoint blockade if tumors express PD-L1. However, chemotherapy does not adequately treat this aggressive, often chemotherapy-resistant subtype of breast cancer. Clinical trials, especially randomized studies are challenging to conduct in MpBC due to the rarity of the diagnosis and the highly aggressive nature, with dismal overall survival measured in months. The estimated overall response rate (ORR) of patients with metastatic MpBC first-line chemotherapy+/- immune checkpoint inhibitors is 15%, beyond the first line under 5%, although the exact ORR is hard to determine accurately given the rarity of this disease. Therefore, combination strategies to understand and overcome mechanisms of resistance to improve marginal chemotherapeutic efficacy are needed to improve the prognosis of these patients. We propose a new two-sample test for a two-sample location problem based on empirical distribution function which is based on rank-order to not only detect the differences between the control and treatment arm but also save time and cost in clinical trials using auxiliary information. The test statistic is constructed as a power divergence between empirical

distribution functions obtained from the two independent samples making it more powerful than its competitors under heavy-tailed and light-tailed distributions. The permutation principle is used to implement the test. We show that our test is component-wise scale invariant. Further, the distribution of the proposed test statistic is obtained under the null hypothesis and in general. We also report the theoretical expectation and variance of the proposed test statistic when the null hypothesis is true. Using the Monte Carlo method we computed empirical power which shows that our test performs better than its competitors under heavy-tailed, light-tailed, and even elliptically asymmetric population distribution. Overall, the proposed test provides better power than its competitors considered here irrespective of the nature of the population.

Binary Randomized Response Technique (RRT) Models Under Measurement Error

William McCance
UC Santa Barbara
williammccance@ucsb.edu
Coauthors: Sat Gupta, University of North Carolina Greensboro; Sadia Khalil, Lahore College for Women University Lahore, Pakistan; Wenhao Shou, University of North Carolina Greensboro

In real-world surveys, measurement error is inevitable as the difference between the actual value of the variable being measured and its recorded value. Many authors in the field of Randomized Response Technique (RRT) have studied the impact of measurement error on quantitative RRT models, but there are no studies founded on binary RRT models. In this article, we propose a binary RRT model under measurement error based on the previous work of Warner (1965) and also introduce the modulo operation to define the measurement error. A simulation study is presented to validate the theoretical findings. Simulations show that the measurement error factor cannot be ignored when using binary RRT models, and the proposed estimator from the binary RRT model under measurement error performs well.

Time-to-ICU Transfer by Frailty Severity using Electronic Health Record Data: Does Nursing Flowsheet Data Matter?

Thomas P. McCoy and Marjorie Jenkins
Department of Family & Community Nursing, School of Nursing, UNC Greensboro; Office of Nursing Research, Cone Health
tpmccoy@uncg.edu

There is a paucity of knowledge about factors affecting ICU transfer of hospitalized patients who do not first receive intensive care upon admission. Among the possible factors is frailty, a clinical syndrome resulting from age-related declines, multisystem physiologic dysregulation, and failed integrative responses to stressors. One potential reason for the sparsity of available related studies is the difficulty in extracting and managing such unit transfer data within the EHR. This presentation will discuss challenges related to the data collection and management for these constructs from the EHR, and how frailty severity affects time-to-ICU transfer based on Cox modeling after missing data adjustment. Frailty risk scores both incorporating nursing flowsheet data elements and without are compared for their utility in exploring ICU transfer.

Benfordness of Measurements Resulting From Box Fragmentation

Zoe McDonald and Livia Betti
Boston University; University of Rochester
zoeann123mcd@gmail.com
Coauthors: Irfan Durmić, University of Jyväskylä; Jack Miller, Yale University; Steven J. Miller, Williams College; Santiago Velazquez-Iannuzzelli, University of Pennsylvania

One might expect that in an arbitrary data set, each number from 1 to 9 is equally likely to be a leading digit, occurring roughly 11% of the time; but in fact, for many data sets, the probability of a leading digit being d base 10 follows the relation $\log_{10}(1+1/d)$, meaning the probability of a leading 1 is about 30%. This phenomenon is known as Benford's law. Motivated by its application to nuclear fragmentation, Benford's law has recently been used to study physical fragmentation processes. We consider the fragmentation of an m-dimensional box and study the d-dimensional measurements of the resulting m-dimensional sub-boxes. Relying on special properties of the Mellin transform, techniques from Fourier analysis, and a maximum-side argument utilizing the Berry-Esseen Theorem, we prove that for $1 \leq d \leq m$, the d-dimensional measurements of the m-dimensional sub-boxes follow Benford's law.

DiseaseNet: a Unified Approach to Disease Detection

Bailey Meche
University of Louisiana at Lafayette
bailey.meche@gmail.com
Coauthors: Benjamin Ginnett, Eastern Arizona College; Kelly Zhou, University of North Texas; Steven Gore, University of North Texas; Rajeev Azad, University of North Texas

Many diseases are correlated to detectable methylation data biomarkers. Methylation signatures of noninfectious diseases could reveal disease presence before symptoms occur, allowing for earlier disease surveillance and intervention. DiseaseNet seeks to provide a unified approach to disease detection using methylation by training a deep learning model on cancer, schizophrenia, asthma, arthritis, and normal samples. Using transfer learning, we trained an additional VAE classifier onto the CancerNet model [1] to classify the original 33 cancer types, 3 new disease types, and one normal. Training this model resulted in an organized latent space and an average F-measure of approximately 89.2%. This reveals not only that general disease signatures exist in methylation data, but also that these diseases, cancer, and normal samples are distinguishable from one another. This model serves as a proof of concept that many human diseases may be encodable into the same model.

Nurses' Experiences Caring for Patients with Opioid Use Disorders

April W. Messer
School of Nursing, College of Health and Human Sciences, Western Carolina University
awmesser@email.wcu.edu

The purpose of this study was to examine nurses' experience of quality care for hospitalized patients with a history of opioid use disorder or self-injection of opioids and whether hospital, unit, or nurse characteristics impacted experiences of quality care. A secondary aim of the study was to understand how nurses' experiences of quality care for this population have been impacted by the COVID-19 pandemic. A national sample of 179 nurses completed an online survey regarding their experiences caring for patients with opioid use disorders. Findings will be discussed on how they contribute to current evidence regarding the need for standardized hospital policies and practices aimed at improving quality care for patients with opioid use disorders.

Identifying Subgroups of Adolescents With Depression Suicidal Ideation: A Look at the TX-YDSRN Data

Abu Minhajuddin
University of Texas Southwestern
zhilam0901@gmail.com
Coauthors: Holli Slater, Manish K Jha, Madhukar H Trivedi

Major depressive disorder (MDD) affects one in six youth and is a leading cause of deaths due to suicide. Furthermore, suicide is the second leading cause of death in youths, and the annual rates of deaths due to suicide have increased by more than 50% increase over the past two decades. Therefore, depression and suicide in youth are critical public health problems in the United States , including the state of Texas. A key knowledge gap in developing targeted treatments is that depression in youths is multi-dimensional in nature. To address this knowledge gap, the state of Texas has funded the Youth Depression and Suicide Network (TX-YDSRN), a state-wide research consortium that has recruited over one thousand youths, ages 8-20 years, in treatment for and/or experiencing depression or suicidal ideation. These youths (and their parents/guardians when applicable) provided data using an electronic data capture system via validated self-reports and clinical/assessor rated instruments. In the present report, we analyzed data collected using pediatric Patient Reported Outcomes Measurement Information System (PROMIS-25) from n = 657 adolescents (age 12-17 years) at their baseline visit (study entry). The goal of the analysis was to classify youth participants based on the PROMIS domains of physical function, depression, anxiety, fatigue, peer relationships, pain interference, and pain severity. The analysis revealed 4 latent classes of participants. Participants in the three of the four groups had higher levels of depression, anxiety, and fatigue while participants in the fourth group were highly functional and had lower levels of depression, anxiety, fatigue, and lower pain interference as well as pain severity. Interestingly, one of the high symptom groups had levels of functionality that were similar to or better than the fourth group (highly functional with low symptom levels) suggesting the presence of factors that may promote resilience and overall well-being even in the presence of significant clinical symptom burden. Participants in the four groups were then compared in terms of socio-demographic characteristics as well as clinical measurements including suicidal ideation, resilience, exposure to traumatic events, and comorbid medical problems.

A New Look at the Search Design Concept

John Morgan
Department of Statistics, Virginia Tech
jpmorgan@vt.edu

In a publication that anticipated future developments in model robustness and discrimination, J. N. Srivastav (1975) introduced the notion of search designs for factorial experiments. He showed how search designs impart an identifiability property relative to all models under consideration in a given, noiseless experimental setting. Here an alternative proof is provided for what has come to be known as the fundamental theorem of the search linear model, leading to several generalizations. The noiseless condition is replaced by a more readily interpretable condition requiring there be no model aliasing, that is, that the competing models partition the mean response space. Implications of the generalizations for design selection are discussed.

Initiatives of Emergency Response and COVID-19 Pandemic Preparedness for Health System Strengthening to Combat COVID -19 Pandemic in Bangladesh

Shah Golam Nabi
DGHS, Ministry of Health and Family Welfare of Bangladesh
sgnabi5@gmail.com
Coauthor: Aminur Rahman

Background: On the eve of the New Year 2020, Chinese health authorities reported an unknown cause of pneumonia nod death related to this disease in the Province of Wuhan in China. As the world was busy celebrating 2020. Very few people give a little thought to this. Within days that unknown disease would halt the world and bring down the whole mankind to its knee. On January 30 World Health Organization declared the disease as Public Health Emergency as International Concern and on 12 March, 2020 it was declared as a Pandemic. Initially named as Wuhan Virus due to its first reporting, WHO officially named it COVID-19. On March 16, 2020, the Government closed all educational institutes (schools, colleges and Universities) to contain the COVID-19 outbreak. On 18 March 2020, the Government of Bangladesh adopted the National Preparedness and Response Plan (NPRP) For COVID-19 with a total cost of USO 29,550,000 million. In Bangladesh, from 3 January 2020 to 12:47 pm CEST, 15 July 2021, there have been 1,059,538 confirmed cases of COVID-19 with 17,052 deaths, reported to WHO (see here). To combat the COVID-19 situation, MOHFW, Bangladesh has taken two projects and many other specific activities. COVID-19 Emergency response and Pandemic Preparedness is one of the ever big projects aimed to support emergency response and to build systems for preparedness to combat COVID-19 as well as any pandemic coming forward.

Goal of the project: Enhancing the capacity of Bangladesh for responding to COVID-19 pandemic and strengthening infectious diseases prevention, control, treatment capacity and emergency preparedness.

Objectives of the Project: To strengthen health systems and health workforce preparedness for Emergency COVID-19 Response by Case Detection, Confirmation, Contract Tracing, Recording and Reporting; Social Distancing Measures; Health System Strengthening and Communication Preparedness; To strengthen and scale up National and Sub-national Health System for Health Emergency Prevention and Preparedness by strengthening Emergency Operation Centre (EOC); increase District Level Surveillance Capacity; and, Stockpiling of critical medical supplies.

Interventions/Input: Procurement of vaccine and extensive vaccination program nation wide; Establishment of vaccine Testing Laboratory at central Level; Establishment of Medical Centres at point of entry (3 air ports and 2 sea ports) with screening facility for Case Detection, Confirmation, Contact Tracing, Recording, Reporting; Establishment of another 27 modern microbiology lab with rt PCR test facility to detect COVID cases; Strengthening of Contact Tracing and Active case identification nationwide through domiciliary services; Mobilizing the local community (trusted community groups, community leaders, local networks and media personnel) to support massaging for social distancing; Establishment of 540 intensive care beds and 860 isolation beds at different facility nationwide; Enhanced Hospital Capacity to Strengthening in patient care and surge capacity (demarked area for flu syndrome, procured and supply of required logistics) Renovation; Establishment of Liquid Medical Oxygen system and ensured uninterrupted oxygen supply at 30 health facilities nation wide ; Establishment of IPC, IDU and epidemiological unit at 64 district health administration office; Mitigation Risk communications Through TVC, Corona BD APP development; Strengthening of Public health emergency centers & surveillance capacity; Capacity building of health care providers, health managers and policy makers different area (contact tracing, surveillance, etc; Conducting research /operational research in different thematic area.

Output: Total Lab Test: 14538033 person, confirm case 2000279, Recovered 1931494, death due to covid -19 29256, vaccination 2nd dose 120238824. Total registration 119221953 (84% reg against population), AEFI was 1183, Total covid dedicated bed 39940 and High flow nasal canula is used during covid 19 was 4828/1318. Total ventilator 2584 where Total no of oxygen concentrator 8540. Total no fulled Oxygen cylinder with 89078, Total no of bed with central oxygen supply 41244; Covid 19 management guidelines for care Giver and policy makers; Well equip infrastructure to optimize the care for infected patients with Covid-19; Prevention of human to human transmission including secondary infections among close contact; Real time community-based disease surveillance; Management information systems for Covid -19 Report; Community Awareness.

Recommendation: Future pandemic management health care organizations should have emphasized communication, connection, and innovation which can effectively address the challenges to adjust capacity, re-

design care models, redeploy staff and increase budget allocation.

Hypergraph Co-Optimal Transport
Tom Needham
Florida State University
tneedham@fsu.edu
Coauthors: Samir Chowdhury, Stanford University; Ethan Semrad, Florida State University; Bei Wang, University of Utah; Youjia Zhou, University of Utah

Hypergraphs capture multi-way relationships in data, and they have consequently seen a number of applications in higher-order network analysis, computer vision, geometry processing, and machine learning. This talk will describe theoretical foundations for studying the space of hypergraphs using ingredients from optimal transport. We will introduce a hypergraph distance based on the co-optimal transport framework of Redko et al. and study its theoretical properties. We will also describe metric and category-theoretic properties of various methods for simplifying hypergraphs, which have so far only been considered as ad hoc constructions.

Balancing Wind Energy Production and Bat Fatalities
Leslie New
Ursinus College
lnew@ursinus.edu
Coauthors: Carl Donovan and Rodrigo Wiff, DMP Stats

Renewable energy is often considered "green" because of the lack of greenhouse gasses associated with energy production. However, the facilities themselves require alterations of the landscapes in which they are placed, and their operation can negatively impact wildlife. Wind facilities, both terrestrial and offshore, are arguably the most common source of renewable energy, but are known to interact negatively with bird and bat species through collisions which led to the injury and death. A common mitigation tactic is known as curtailment, when the operation of the turbines is halted or delayed to avoid time periods and/or environmental conditions that are more likely to lead to wildlife fatalities. While effective, curtailing turbines also reduces energy generation, creating a tradeoff between the risk posed by the operating turbines and energy production. Improving our understanding of the factors contributing to species' collision risk can enable curtailment to be used more effectively, avoiding fatalities while increasing energy generation. With that in mind, we built a bat fatality risk model for a terrestrial wind facility in North America, using data on weather, bat acoustic activity, thermal imaging and fatalities in 2021. A two-stage process was taken, first modelling bat activity as a function of time and weather conditions, and then modelling bat fatalities as a function of activity. Different models were considered, such as gradient boosted machines, but a Generalized Additive Mixed Model was found to be most appropriate for activity, while a simple proportional model was required for bat fatalities due to the sparsity of the available data. These modeled relationships then provided the basis for the consideration of curtailment rules in which the conditions of turbine operation were objectively measured against implied fatality rates and power outputs to determine the best balance. After optimization, we found the potential for an additional 8.5% reduction in bat fatalities relative to blanket curtailment (all turbines non-operating below wind speeds of 5 m/s), with no additional power loss, by using only slightly more complex rules. A potential 10% reduction in bat fatalities was possible compared to blanket curtailment if curtailment rules vary by month, again with no additional power loss.

Monitoring Sequential Structural Changes in Penalized High-Dimensional Linear Models

Wei Ning
Bowling Green State University
wning@bgsu.edu
Coauthor: Suthakaran Ratnasingam, Department of Mathematics, California State University, San Bernardino

In this talk, we propose a procedure to monitor the structural changes in the penalized regression model for high-dimensional data sequentially. Our approach utilizes a given historical data set to perform both variable selection and estimation simultaneously. The asymptotic properties of the test statistics are established under the null and alternative hypotheses. The finite sample behavior of the monitoring procedure is investigated with simulation studies. The proposed method is applied to a real data set to illustrate the detection procedure.

Subdata Selection and TreeS

Rhys O'Higgins
Macalester College
rohiggin@macalester.edu
Coauthors: John Stufken, George Mason University; Rakhi Singh, Binghamton University; Samuel Griffin, University of North Carolina, Asheville)

The sheer size of modern datasets often yields making predictions from the full data computationally infeasible or impossible. To circumvent this problem, smaller subdata is selected from the original dataset, and used to make predictions and analyses. We propose a new subdata selection method, TreeS, aimed at maximizing predictive accuracy garnered from the subdata, that combines elements from existing methods Supersompress (Joseph and Mak, 2020) and Twinning (Vakayil and Joseph, 2021). Early results suggest TreeS is a strong alternative to existing methods, both in quality of results and algorithmic efficiency.

RCD and TDA for 2D Scenes Extracted From Electronic Images

Vic Patrangenaru
Florida State University
vic@stat.fsu.edu
Coauthor: Robert L. Paige, Missouri University of Science and Technology

We consider statistical and topological data analysis of 2D image data. In the first part of this talk, we consider methodologies based on the Region Covariance Descriptor (RCD). The second part is dedicated to Topological Data Analysis (TDA) based on simplicial and cubical persistent homologies. Our 2D example concerns images of leaf data from Qiu et al. (2019), which consist of pictures of two leaves-A and B- from the same tree, twenty of each leaf, from different perspectives. These novel statistical procedures are used for correctly determining that leaf A images and leaf B images are in fact those of different leaves and also for correctly classifying new out-of-sample leaf images.

References

Qiu, Mingfei; Paige, Robert; Patrangenaru, Vic (2019). A nonparametric approach to 3D shape analysis from digital camera images—II. J. Appl. Stat. 46, no. 15, 2677–2699"

Inference for Distance-to-Set Regularization via Constrained Bayesian Inference
Rick Presman
Duke University
rick.presman@duke.edu
Coauthor: Jason Xu

A fundamental consideration for Bayesian modeling is how to deal with constraints. These can be particularly challenging to address if they are exogenously specified, or superimposed on the model and priors. Inspired by the majorization-minimization (MM) literature, we propose a class of priors, which we call distance-to-set priors, that act as mechanism for constraint relaxation. We draw connections between the world of Bayesian modeling and optimization, and we show that this class of priors has desirable theoretical properties in relation to constrained Bayesian modeling. Furthermore, we elucidate why distance-to-set priors are particularly amenable to gradient-based sampling algorithms and can succeed in sampling the posterior in situations in which one is limited in the ability to relax the constraints. Finally, we demonstrate our results in various simulated and real-world settings.

Coarse Embeddability of the Space of Persistence Diagrams and Wasserstein Space
Neil Pritchard
UNC Greensboro
cnpritch@uncg.edu
Coauthor: Thomas Weighill

When applying machine learning and statistical techniques one implicitly requires that they are working in a Hilbert space. When working in metric spaces not naturally contained in Hilbert space one may then seek a mapping into Hilbert space with controlled distortion. One such family of mappings are called coarse embeddings. It remains an interesting open question whether the space of persistence diagrams and Wasserstein space on R^2 coarsely embed into Hilbert space for p¡2. In talk I will examine the connection between these problems and show they are equivalent for certain values of p.

Embedding Functional Data: Multidimensional Scaling and Manifold Learning
Wanli Qiao
George Mason University
wqiao@gmu.edu
Coauthor: Ery Arias-Castro, University of California, San Diego

We adapt concepts, methodology, and theory originally developed in the areas of multidimensional scaling and dimensionality reduction for multivariate data to the functional setting. We focus on classical scaling and Isomap — prototypical methods that have played important roles in these areas — and showcase their use in the context of functional data analysis. In the process, we highlight the crucial role that the ambient metric plays.

Joint work with Ery Arias-Castro (University of California, San Diego).

How Auxiliary Information Can Help Your Missing Data Problem

Jerry Reiter
Duke University
jreiter@duke.edu

Many surveys (and other types of databases) suffer from unit and item nonresponse. Typical practice accounts for unit nonresponse by inflating respondents' survey weights, and accounts for item nonresponse using some form of imputation. Most methods implicitly treat both sources of nonresponse as missing at random. Sometimes, however, one knows information about the marginal distributions of some of the variables subject to missingness. In this talk, I discuss how such information can be leveraged to handle nonignorable missing data, including allowing different mechanisms for unit and item nonresponse (e.g., nonignorable unit nonresponse and ignorable item nonresponse).

Markov Chain Composite Likelihood and Its Application in Genetic Recombination Model

Grace Rhodes
Mount Holyoke College Department of Mathematics & Statistics (2022), Duke University School of Medicine Department of Biostatistics & Bioinformatics (Ph.D. Student)
rhode22g@mtholyoke.edu
Coauthor: Jianping Sun, Department of Mathematics & Statistics, University of North Carolina at Greensboro

DNA sequencing technologies are rapidly advancing, allowing researchers access to data which is both high quality and highly detailed, especially on identifying single nucleotide polymorphism (SNP) sites on individual haplotypes. These SNPs play an essential role in enriching our understanding of human evolutionary history through hierarchical trees of SNP inheritance and to identify SNPs associated with disease. However, while this detailed SNP data exists for extant humans, this is not the case for their ancestors. There is a need to estimate backward across generations. Approached statistically, the question is: If we observe current descendants' SNP sequences, how can we estimate the unknown ancestor's SNP sequence while considering biological complexities (such as mutation and recombination)? Previously, Chen and Lindsay (2006) proposed the ancestor mixture model to construct the hierarchical trees. However, this model only focused on mutation and without taking recombination into account.

The present study proposes a Recombination Model, which estimates the unknown ancestral distribution of SNP sequences from observed descendant sequences while considering a fixed probability of recombination. While a log-likelihood statement for this model can be explicitly obtained, the large number of recombination possibilities in a sequence and the complex parameter space make a traditional MLE approach computationally infeasible. We instead use Markov Chain Composite Likelihood (MCCL) and hierarchical estimation to estimate the unknown ancestral distribution. Simulation studies show that the estimator performs well in obtaining estimates of the marginal and joint distributions; bias and variance were minimal. However, we observe a tendency towards underestimating joint probabilities, a directional effect of the Markov chain, and a bias-variance tradeoff relative to the Markov chain order.

Linear Models for Doubly Multivariate data with Exchangeably Distributed Errors

Anuradha Roy
The University of Texas at San Antonio
Anuradha.Roy@utsa.edu
Coauthor: Timothy Opheim

Doubly multivariate data, where observations are made on p response variables and each response variable is measured over n sites or time points, construct matrix-valued response variable, and arise across a wide range of disciplines, including medical, environmental and agricultural studies. The popularity of the classical general linear model (CGLM) is mostly due to the ease of modeling and authentication of the appropriateness

of the model. However, CGLM is not appropriate and thus not applicable for correlated doubly multivariate data. We propose an extension of linear model for doubly multivariate data with exchangeably distributed errors for multiple observations. Maximum likelihood estimates of the matrix parameters of the intercept, slope and the eigenblocks of the exchangeable error matrix are derived. The practical implications of the methodological aspects of the proposed extended model for doubly multivariate data are demonstrated using two medical datasets.

A Comparison of Multiple Testing Procedures for Controlling the False Discovery Rate Under Unequal Variances

Fatema Ruhi
Southeast Missouri State University
fatemaruhi131209@gmail.com
Coauthor: Pradeep Singh, Department Of Mathematics, Southeast Missouri State University

The multiple hypothesis testing is used to control the False Discovery Rate (FDR) in genomic studies. The FDR is the expected proportion of the number of the False Positives from all the rejected null hypotheses. The most commonly used multiple testing procedures were given by Storey - Tibshirani (2003) and Benjamini - Hochberg (1995). Storey's q value is similar to p value and is based on FDR rather than False Positive Rate. The q value method uses the p values obtained from permutation test for equality of two means. Li et. al. (2017) proposed a new method based on Bonferroni's approach and named it Bon-EV procedure. Neubert and Brunner (2007) proposed a non-parametric method to address the Behrens-Fisher problem of unequal variances. This study proposes a modification of Benjamini - Hochberg, Bon-EV, and Storey's q value methods to control the FDR under the assumption of unequal variances. The modified procedures will use the p values obtained by using Neubert and Brunner non-parametric test. Monte Carlo method will be used to compare the above testing procedures for their ability to control FDR to its nominal level for populations with unequal variances.

A Bayesian Analysis of Two-Stage Randomized Experiments in the Presence of Interference, Treatment Nonadherence, and Missing Outcomes

Arman Sabbaghi
Purdue University
sabbaghi@purdue.edu
Coauthor: Yuki Ohnishi, Department of Statistics, Purdue University

Three critical issues for causal inference that often occur in modern, complicated experiments are interference, treatment nonadherence, and missing outcomes. A great deal of research efforts has been dedicated to developing causal inferential methodologies that address these issues separately. However, methodologies that can address these issues simultaneously are lacking. We propose a Bayesian causal inference methodology to address this gap. Our methodology extends existing causal frameworks and methods, specifically, two-staged randomized experiments and the principal stratification framework. In contrast to existing methods that invoke strong structural assumptions to identify principal causal effects, our Bayesian approach uses flexible distributional models that can accommodate the complexities of interference and missing outcomes, and that ensure that principal causal effects are weakly identifiable. We illustrate our methodology via simulation studies and a re-analysis of real-life data from an evaluation of India's National Health Insurance Program. Our methodology enables us to identify new significant causal effects that were not identified in past analyses. Ultimately, our simulation studies and case study demonstrate how our methodology can yield more informative analyses in modern experiments with interference, treatment nonadherence, missing outcomes, and complicated outcome generation mechanisms.

Accounting for Lack of Trust in Optional Binary RRT Models Using a Unified Measure of Privacy and Efficiency

Pujita Sapra
UNC Greensboro
p_sapra@uncg.edu
Coauthors: Sadia Khalil, Department of Statistics, LCWU, Lahore, Pakistan; Sat Gupta, University of North Carolina Greensboro

In this study, we firstly present a non-optional mixture of the Warner and Greenberg binary models proposed by Lovig et al. [2021]. This model accounts for untruthful responses due to a lack of respondent trust in the traditional binary RRT models. In this work, we examine an optional version of the Lovig et al. [2021] model and show, theoretically and empirically, that it works better than the corresponding non-optional model.

Active Learning for Deep Gaussian Process Surrogates

Annie Sauer
Department of Statistics, Virginia Tech
anniees@vt.edu
Coauthors: Robert B. Gramacy, David Higdon

Deep Gaussian processes (DGPs) are increasingly popular as predictive models in machine learning for their non-stationary flexibility and ability to cope with abrupt regime changes in training data. Here we explore DGPs as surrogates for computer simulation experiments whose response surfaces exhibit similar characteristics. In particular, we transport a DGP's automatic warping of the input space and full uncertainty quantification, via a novel elliptical slice sampling Bayesian posterior inferential scheme, through to active learning strategies that distribute runs non-uniformly in the input space – something an ordinary (stationary) GP could not do. Building up the design sequentially in this way allows smaller training sets, limiting both expensive evaluation of the simulator code and mitigating cubic costs of DGP inference. When training data sizes are kept small through careful acquisition, and with parsimonious layout of latent layers, the framework can be both effective and computationally tractable. Our methods are illustrated on simulation data and two real computer experiments of varying input dimensionality. We provide an open source implementation in the "deepgp" package on CRAN.

Mapper-Type Algorithms: Extensions and Generalizations

Radmila Sazdanovic
North Carolina State University
rsazdan@ncsu.edu
Coauthors: Pawel Dlotko and Davide Gurnari, DIOSCURI Centre in Topological Data Analysis, Warsaw, Poland

Mapper and Ball Mapper are Topological Data Analysis tools used for exploring high dimensional point clouds and visualizing scalar–valued functions on those point clouds. Inspired by open questions in knot theory, new features are added to the Ball Mapper that enable encoding the structure, relations within and symmetries of the point cloud. New hybrid Mapper–on–Ball Mapper algorithm that extends Mapper algorithm from 1-dimensional to lens functions whose range is high dimensional is introduced. Moreover, the strengths of Mapper and Ball Mapper constructions are combined to create a tool for comparing high dimensional data descriptors of a single data set. As a proof of concept we include applications to knot and game theory, as well as material science. Finally, we compare our results describing the structure of polynomial knot invariants with those obtained using statistical and ML tools.

Performance of Optional Unrelated Question Randomized Response Models under Two-Stage Stratified Cluster Sampling for Estimation of Population Proportion and Sensitivity Level

Javid Shabbir
University of Wah, Wah Cantt, Pakistan
javidshabbir@gmail.com
Coauthor: Sat Gupta, University of North Carolina at Greensboro, US

Optional unrelated question randomized response models by Narjis and Shabbir (2020) are extended to stratified two-stage cluster sampling. The basic premise of optional RRT model, that a question may be sensitive for one respondent but may not be sensitive for another. In an optional RRT model, a respondent is requested to provide a scrambled response only if he/she considers the question is sensitive. Otherwise, the respondent provides a truthful response. Two-question approach is extended to unequal probability sampling for parameters estimation, the prevalence sensitive characteristics and the sensitivity level. The extended models are more efficient than the competitor models for a population proportion having a sensitive attribute.

Predictors of Depression and Anxiety Among Urban Population During COVID-19: An Online Cross-Sectional Survey

MD Shahjahan
Daffodil International University
mdshahjahan@agnionline.com
Coauthors: Md. Nawal Sarwer and Nadira Mehriban, Daffodil International University

The novel coronavirus (COVID-19) pandemic has emerged as a major public health crisis, which not only threatens the lives of people but also affects their mental health and well-being. Bangladesh is facing enormous challenges with COVID-19 pandemic as other countries around the world. This pandemic challenged both the physical and mental health of the people from all segments of the society. This study explored the depression and anxiety level among the adults living in Dhaka City.

Data were collected online during Mar-May 2021 following the shorter version of Depression, Anxiety and Stress scale (DASS21). The self-administered electronic questionnaire covered socio-demographic information, history of COVID-19, physical and mental health condition of the respondents during pandemic, etc. Respondents were asked to indicate the level of symptoms present on a 4-point Likert scale ranging $0-3$. The respondents were informed of the study purpose, ethical electronic consent was obtained and they were assured that all information provided by them kept confidential and anonymous. Data were analyzed as per standard practice.

The web-based cross-sectional survey covered 981 respondents were used for analysis. Among the respondents, almost half were male that corresponds to the national male-female ratio. The mean age of the respondents were 29.4 ($SD \pm 9.3$) years and nearly three-fourth was young adults of age less than 30 years. More than 65% of the respondents had bachelor degree or above and nearly a quarter had secondary level of education. Respondents enrolled from various professional backgrounds where, nearly half were healthcare providers, one-fourth involved in business and one-fifth were students. The average monthly income per household was reported to be BDT 40545 (1 USD = 86 BDT). Of the respondents, one-third reported that they themselves or their family members or near-relatives were suffered from COVID-19.

Slightly over one-fourth of the respondents had no anxiety and 22% had no depression. Among the respondents with anxiety, 35% had mild, 31% had moderate and 9% had severe level of anxiety. While for the respondents with depression, 32% were mild, 26% were moderate and 20% were of severe level. The level of education ($p < 0.023$) and asset quintiles ($p < 0.026$) were found to be significant predictors for anxiety. On the other hand, level of education ($p < 0.001$), occupation ($p < 0.005$) and asset quintiles ($p < 0.005$) were significant predictors of depression.

Bangladesh is experiencing demographic dividend period. If the active young population suffered from mental health problems, it will hinder the economic development of the family as well as of the country. In case of such pandemic, the mental health counseling is equally important in addition to existing healthcare

system. However, Bangladesh is lacking of mental health counselors/clinical psychologists. Policymakers need to take appropriate measures to address the problem in near future.

Confidence Band Approach for Comparison of COVID-19 Case Counts
Qin Shao
University of Toledo
qin.shao@utoledo.edu

The outbreak of the novel coronavirus was declared to be a global emergency in January, 2020, and everyday life throughout the world was disrupted. Among many questions about COVID-19 that remain unanswered, it is of interest for society to identify whether there is any significant difference in daily case counts between males and females. The daily case count sequences are correlated due by the very nature of the contagious disease caused by the novel coronavirus, and contain a nonlinear trend due to several unexpected events, such as vaccinations and the appearance of the delta variant. It is possible that these unexpected events have changed the dynamic system that generates data. The classic t-test is not appropriate to analyze such correlated data with a nonconstant trend. This study applies a simultaneous confidence band approach in an attempt to overcome these challenges; that is, a simultaneous confidence band for the trend of an autogressive moving-average time series is constructed using a B-splines estimation. The proposed method is applied to the daily case count data of male and female seniors (at least 60 years old) in the State of Ohio from April 1, 2020 to March 31, 2022, and the result shows that there is a significant difference at the 95% significance level between the two gender case counts adjusted for the population sizes.

Semi-Supervised TDA
Don Sheehy
North Carolina State University
drsheehy@ncsu.edu
Coauthor: Kirk Gardner

One of the most pervasive metaphors in TDA is that homology generalizes clustering. This situates techniques like persistent homology firmly in the domain of unsupervised learning. Although less common, there is also substantial work on the estimation of a persistence barcode of an unknown function from samples—a kind of supervised TDA problem. In this talk, I will consider the question of what kinds of guarantees are possible if one has evaluations of the function at only a subset of the input points—a semi-supervised TDA problem. I will summarize the new theory of sub-barcodes and show how one can compute a barcode that is guaranteed to be contained in the barcode of every Lipschitz function that agrees with the sample data. This provides strong guarantees even with only partial data.

ESPs: A New Cost-Efficient Sampler for Expensive Posterior Distributions

Flora Shi
Duke University
cs558@duke.edu
Coauthors: Irene Ji, Simon Mak

A key computational bottleneck for Bayesian inverse problems is the expensive evaluation cost of forward simulation models, which can require thousands of CPU hours for simulating complex physical processes. While state-of-the-art posterior sampling algorithms (e.g., Hamiltonian Monte Carlo methods) may be "sample-efficient", i.e., they provide a good representation of the posterior given limited samples, such methods can be highly cost-inefficient, since they require at least one evaluation of the forward model per sample. Given a fixed computational budget, we wish to have a "cost-efficient" posterior sampler which yields a good representation of the desired posterior given limited forward model evaluations. We thus present a new sampler called cost-Efficient Stein Points (ESPs) for this goal. Our ESPs extend the recently-proposed Stein points in Chen et al. (2018, ICML), which makes use of a sequential minimization of the kernel Stein discrepancy (KSD) for sample-efficient posterior exploration. The key novelty of ESPs is the use of a carefully-constructed Gaussian process surrogate model of the KSD, which allows for cost-efficient sequential minimization via Bayesian optimization. We demonstrate the cost-efficiency of ESPs over state-of-the-art posterior sampling algorithms via a suite of numerical experiments and a calibration application for a falling object subject to drag.

A Projection Space-Filling Criterion and Related Optimality Results

Chenlu Shi
Colorado State University
chenlu.shi@colostate.edu
Coauthor: Hongquan Xu, University of California, Los Angeles

Computer experiments call for space-filling designs. Recently, a minimum aberration type space-filling criterion was proposed to rank and assess a family of space-filling designs including Latin hypercubes and strong orthogonal arrays. It aims at capturing space-filling properties of a design when projected regarding the volume of grids instead of the dimension. However, in this paper, we take both the volume of grids and the dimension into account by proposing first an expanded space-filling hierarchy principle and then a projection space-filling criterion as per the new principle. When projected onto grids of the specific volume, the proposed criterion ranks designs via sequentially maximizing the space-filling properties on equal-volume grids in lower dimensions to higher dimensions, while the minimum aberration type space-filling criterion compares designs by maximizing the aggregate space-filling properties on multidimensional grids of the same volume. We present illustrative examples to tell two criteria and conduct simulations to demonstrate the utility of our criterion in terms of selecting efficient space-filling designs to build statistical surrogate models. We further consider the construction of the optimal space-filling designs under the proposed criterion. Although many algorithms have been proposed for generating space-filling designs, it is well-known that they often deteriorate rapidly in performance for large designs. In the present paper, we develop some theoretical optimality results and characterize several classes of strong orthogonal arrays of strength three that are the most space-filling.

Kernel Density Estimation Using Additive Randomized Response Technique (RRT) Models
Wenhao (Wendy) Shou
UNC Greensboro
w_shou@uncg.edu
Coauthor: Sat Gupta

In 2002 Ahmad introduced the kernel estimation of the density curve of a sensitive variable based on multiplicative RRT models and provided some theoretical results. In this article, we propose a kernel density estimator in the context of additive RRT models, which are more commonly used in the field of survey sampling. A simulation study is presented to validate the theoretical results from the previous work of Ahmad, and also compare the performances of the density estimators based on the additive and multiplicative RRT models. Simulations show that the proposed density estimator using additive scrambling performs better than the one using multiplicative scrambling, and it allows more error in the bandwidth selection in kernel density estimation.

Analytical Tools for Whole-Brain Networks: Fusing Statistics and Network Science to Understand Brain Function
Sean L. Simpson
Wake Forest University School of Medicine
slsimpso@wakehealth.edu

Brain network analyses have exploded in recent years, and hold great potential in helping us understand normal and abnormal brain function. Network science approaches have facilitated these analyses and our understanding of how the brain is structurally and functionally organized. However, the development of statistical methods that allow relating this organization to health outcomes has lagged behind. We have attempted to address this need by developing analytical tools that allow relating system-level properties of brain networks to outcomes of interest. These tools serve as synergistic fusions of statistical approaches with network science methods, providing needed analytic foundations for whole-brain network data. Here we delineate two recent approaches—a mixed-modeling framework for dynamic network analysis and a regression framework for relating distances between brain network features to covariates of interest—that expand the suite of analytical tools for whole-brain networks and aid in providing complementary insight into brain function.

Design Selection for Supersaturated Designs
Rakhi Singh
Binghamton University
agrakhi@gmail.com
Coauthor: John Stufken, George Mason University

An extensive literature is available on design selection criteria and analysis techniques for two-level supersaturated designs. The most notable design selection criteria are the popular E(s2)-criterion, UE(s2)-criterion, and more recently, the Var(s+)-criterion, while the most notable analysis technique is the Gauss-Dantzig Selector. It has been observed that while the Gauss-Dantzig Selector is often the preferred analysis technique, differences in the screening performance of different designs are not captured well by any of the common design selection criteria. In addition, none of the criteria have any direct connection to the Gauss-Dantzig Selector. We develop two new design selection criteria inspired by large sample desiderata of the Gauss-Dantzig Selector. Then, using a multi-objective Pareto-based coordinate exchange algorithm, we find Pareto efficient designs. The obtained Pareto efficient designs perform better in about 85% of the considered cases as screening designs than the Var(s+)-optimal designs, especially when the true signs of effects are known. For the remaining 15% of the cases as well as for the unknown effect signs, the Pareto

efficient designs perform at par with the Var(s+)-optimal designs. If time permits, we will also show the corresponding results for three-level supersaturated designs.

Subdata Selection With a Large Number of Variables
John Stufken
George Mason University
jstufken@gmu.edu
Coauthor: Rakhi Singh, Binghamton University

With ever larger datasets, computational challenges have led to a vast literature on using only some of the data (subdata) for estimation or prediction. This raises the question how subdata should be selected from the entire dataset (full data). One possibility is to select the subdata randomly from the full data, but this is typically not the best method. The literature contains various suggestions for better alternatives. Most of these alternatives focus on situations where the number of variables is small to modest. In this presentation we introduce a method that can be used for big data with a large number of variables in the context of linear regression.

Repeated Sampling in EMA Studies: A Discussion Statistical Challenge and Potential Solution
Jianping Sun
UNC Greensboro
j_sun4@uncg.edu
Coauthor: Xianming Tan, UNC Chapel Hill

Ecological momentary assessment (EMA) is a real-life (ecological) and real-time (momentary) data collection method. It involves repeated sampling of subjects' current behaviors and experiences with the aims to minimize recall bias, maximize ecological validity, and allow study of microprocesses that influence behavior in real-world contexts. One challenge in designing EMA systems is deciding the frequency and timing of repeated survey to achieve a balance between minimizing burden and intrusion to participants and maximizing information collected. In this talk, we discuss the statistical challenge in the design of repeated sampling scheme in EMA studies and propose an intelligent Ecological Momentary Assessment (iEMA) method which aims to address the above challenge. This is achieved through advanced machine learning algorithms that use data already collected to evaluate the redundancy of future survey questions. Based on such evaluations, the iEMA method can 'intelligently' decide when to deliver specific questions so as to minimize the total number of questions asked of participants. We demonstrate this iEMA method using data from an existing EMA study and show that we could eliminate 30% of questions if the iEMA method was utilized.

Functional-Input Gaussian Processes with Applications to Inverse Scattering Problems
Chih-Li Sung
Michigan State University
Coauthor: Ying Hung, Rutgers University

Surrogate modeling based on Gaussian processes (GPs) has received increasing attention in the analysis of complex problems in science and engineering. Despite extensive studies on GP modeling, the developments for functional inputs are scarce. Motivated by an inverse scattering problem in which functional inputs representing the support and material properties of the scatterer are involved in the partial differential equations, a new class of kernel functions for functional inputs is introduced for GPs. Based on the proposed GP models, the asymptotic convergence properties of the resulting mean squared prediction errors are derived and the finite sample performance is demonstrated by numerical examples. In the application to inverse scattering, a surrogate model is constructed with functional inputs, which is crucial to recover the reflective index of an inhomogeneous isotropic scattering region of interest for a given far-field pattern.

Bayesian Predictive Decision Synthesis: Betting on Better Models
Emily Tallman
Department of Statistical Science, Duke University
emily.tallman@duke.edu
Coauthor: Mike West

I present a new framework called Bayesian predictive decision synthesis (BPDS) which explicitly integrates decision-analytic outcomes in the evaluation, comparison, and combination of a set of candidate models. BPDS extends recent theoretical and practical advances in both Bayesian predictive synthesis and empirical goal-focused model uncertainty analysis. This extension is enabled by the development of a novel subjective Bayesian perspective on model weighting in predictive decision settings, with theoretical connections to entropic tilting. BPDS provides a flexible framework focused on context-specific score functions which allows decision makers to stress the intended use of models when evaluating predictions. Specific examples come from applied contexts including optimal design for regression prediction and sequential time series forecasting for financial portfolio decisions.

Ad Marketplace Optimization Towards Auto-Bidding
Ryan Tang
Duke University and Reddit, Inc.
ryan.tang@duke.edu
Coauthor: David Banks, Duke University

For many internet companies, targeted online advertising serves most of their revenue. With the leading success of Google Ads and Facebook Ads, many companies are developing their marketplaces, serving as publishers, and providing new platforms for content generation. However, due to the chaotic nature of a complex system arising from the interaction between users and advertisers — the ad marketplace — publishers are forced to delicately make trade-offs between income, advertisers' constraints, ad quality, and user experiences. Here at Reddit, we have a similar challenge. We use a generalized second-price (GSP) mechanism to rank and price each impression and a naive budget pacing controller to smooth the delivery across the day. The algorithm is sub-optimal for advertisers with constrained budgets and publishers' revenue. Hence, we will discuss the intricacies of the issue, detail how we engineered an auto-bidding solution to alleviate the problem through the lens of statistics, and discuss the next step toward a full auto-bidding product.

Unsupervised Multi-task and Transfer Learning on Gaussian Mixture Models

Ye Tian
Department of Statistics, Columbia University
ye.t@columbia.edu
Coauthors: Haolei Weng, Department of Statistics and Probability, Michigan State University; Yang Feng, School of Global Public Health, New York University

Unsupervised learning has been widely used in many real-world applications. One of the simplest and most important unsupervised learning models is the Gaussian mixture model (GMM). In this work, we study the multi-task learning problem on GMMs, where there are several tasks with potentially similar GMM parameters, and we would like to achieve better performance by utilizing similar structures between them. A multi-task GMM learning algorithm is proposed and we derive high-probability upper bounds for both estimation error of GMM parameters and mis-clustering error, which verifies that when the tasks are close to each other, our method can achieve better upper bounds than those from the single-task GMM. Moreover, we derive (nearly) matching lower bounds, showing the optimality of our method in a wide range of regimes. Also, our method is robust to a small fraction of outlier tasks and is extended to transfer learning, where similar theoretical results are derived and presented. Finally, we demonstrate the effectiveness of our algorithms through simulations and a real data analysis. To the best of our knowledge, this is the first work studying multi-task and transfer learning on GMM with theoretical guarantees.

The Polaron Problem

Srinivasa Varadhan
Courant Institute/NYU
varadhan@cims.nyu.edu
Coauthor: Chiranjib Mukherjee, University of Muenster, Germany

In mathematical terms the problem consists of the following. P is the white noise or the process of increments of the three dimensional Brownian Motion. As a process with independent increments it can be restricted to any interval $[-T, T]$. We will continue to denote it by P.

We define a tilted measure $P^{\alpha,T}$ as

$$\frac{dP^{\alpha,T}}{dP} = \frac{1}{Z(\alpha,T)} \exp\left[\alpha \int\int_{-T\le<s<t\le T} \frac{e^{-|t-s|}}{|x(t)-x(s)|} dtds\right]$$

where $Z(\alpha, T)$ is the normalizing constant

$$Z(\alpha,T) = E^P\left[\exp\left[\alpha \int\int_{-T\le<s<t\le T} \frac{e^{-|t-s|}}{|x(t)-x(s)|} dtds\right]\right]$$

The problems are to show that $\frac{\log Z(\alpha,T)}{2T}$ has a limit $Z(\alpha)$ as $T \to \infty$ and $\frac{Z(\alpha)}{\alpha^2}$ has limit as $\alpha \to \infty$. To analyze the behavior of the measure $P^{\alpha,T}$ as $T \to \infty$ and show that it has a limit P^α as $T \to \infty$. Prove a central limit theorem for $x(T)-x(-T)$ under P^α. Finally to show that limiting variance $\sigma^2(\alpha)$ of $\frac{x(T)-x(-T)}{\sqrt{2T}}$ decays like α^{-4} as $\alpha \to \infty$. This problem arises in modeling the movement of a free electron that interacts with the atoms in a crystal. The electron is slowed down and its effective mass is expected to increase like α^4.

Optimal Cut-Point for Disease Incidence With Censored Data

Cuiling Wang
Albert Einstein College of Medicine
cuiling.wang@einsteinmed.edu
Coauthors: Mindy Katz, Carol Derby, Mimi Kim, Richard Lipton

Selection of cut-points for a marker for early detection of disease, i.e., to identify those at high risk of developing the disease within a given time period in the future is crucial in clinical practice and research. Although optimal thresholds based on various criteria can be obtained through time-dependent receiver operating characteristic (ROC) analyses, the properties of the optimal cut- points are rarely studied and consequently not well known. We investigate the properties of the time-dependent optimal cut-points based on various criteria including setting target level for sensitivity or specificity, Youden's index, weighted sum of sensitivity and specificity which includes Youden's index as a special case, and the average overall cost, which takes the financial and health costs of the test results into account. For the Youden's index, weighted sum and average overall cost criteria, we provide formulae to directly estimate the optimal cuts using survival models. The methods are applied to screening for pre-clinical Alzheimer's dementia using a well-established memory test in the Einstein Aging Study. Simulation studies are performed to evaluate the performance of the proposed estimates and compared to those obtained from the time-dependent ROC analysis.

Statistical Inference for Mean Functions of 3D Functional Objects

Guannan Wang
College of William and Mary
gwang01@wm.edu
Coauthors: Yueying Wang, Columbia University Irving Medical Center; Brandon S. Klinedins, University of Washington; Auriel A. Willette, Iowa State University; Li Wang, George Mason University

Functional data analysis has become a powerful tool for statistical analysis of complex objects, such as curves, images, shapes, and manifold-valued data. Among these data objects, recent 2D or 3D images obtained using medical imaging technologies have been attracting researchers' attention. For example, functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) can provide a very detailed characterization of brain activity. In general, 3D complex objects are usually collected within the irregular boundary, whereas most statistical methods have been focused on a regular domain. To address this problem, we model the complex data objects as functional data and propose multivariate spline smoothing based on triangulation for estimating the mean functions of 3D functional objects. The asymptotic properties of the proposed estimator are systematically investigated where consistency and asymptotic normality are established. We also provide a computationally efficient estimation procedure for covariance function and corresponding eigenvalue and eigenfunctions and derive uniform consistency. Motivated by the need for statistical inference for complex functional objects, we present a novel approach for constructing simultaneous confidence corridors to quantify estimation uncertainty. The extension of the procedure to the two-sample case is discussed with numerical experiments and a real-data application using Alzheimer's Disease Neuroimaging Initiative database.

Maximum Sampled Conditional Likelihood for Informative Subsampling
HaiYing Wang
University of Connecticut
haiying.wang@uconn.edu
Coauthor: Jae Kwang Kim, Iowa State University

Subsampling is a computationally effective approach to extract information from massive data sets when computing resources are limited. After a subsample is taken from the full data, most available methods use an inverse probability weighted (IPW) objective function to estimate the model parameters. The IPW estimator does not fully utilize the information in the selected subsample. In this paper, we propose to use the maximum sampled conditional likelihood estimator (MSCLE) based on the sampled data. We established the asymptotic normality of the MSCLE and prove that its asymptotic variance covariance matrix is the smallest among a class of asymptotically unbiased estimators, including the IPW estimator. We further discuss the asymptotic results with the L-optimal subsampling probabilities and illustrate the estimation procedure with generalized linear models. Numerical experiments are provided to evaluate the practical performance of the proposed method.

Semiparametric Estimation of Non-Ingroable Missingness with Refreshment Sample
Jing Wang
University of Illinois at Chicago
jiwang12@uic.edu
Coauthors: Lan Xue, Oregon State University; Jianfei Zheng, Oregon State University; Elena Graetz, University of Illinois at Chicago

Missing data is one of the major methodological problems in longitudinal studies. It not only reduces the sample size, but also can result in biased estimation and inference. It is crucial to correctly understand the missing mechanism and appropriately incorporate it into the estimation and inference procedures. Traditional methods, such as the complete case analysis and imputation methods, are designed to deal with missing data under unverifiable assumptions of MCAR and MAR. Our focus is on the identification and estimation of attrition (missing) parameters under the non-ignorable missingness assumption using the refreshment sample in two-wave panel data. When one is unable to specify the the joint distribution, we propose a semi-parametric method to estimate the attrition parameters by marginal density estimates with the help of two constraints from Hirano(2001) and the additional information provided by the refreshment sample. We derive asymptotic properties of the semi-parametric estimators and illustrate their performance with simulations. Inference based on bootstrapping is proposed and verified through simulations. A real data application are attempted in the Netherlands Mobility Panel and British Household Panel survey.

Balanced Subsampling for Big Data with Categorical Predictors
Lin Wang
Purdue University
linwang@purdue.edu

The dramatic growth of big datasets presents a new challenge to data storage and analysis. Data reduction, or subsampling, that extracts useful information from datasets is a crucial step in big data analysis. I will introduce a balanced subsampling approach for big data with categorical predictors. The merits of the proposed approach are two-fold: (i) it is easy to implement and fast; (ii) the selected subsample allows robust effect estimation and prediction. Theoretical results and extensive numerical results show that the proposed approaches are superior to simple random subsampling. The advantages of the balanced subsampling approach are also illustrated through the analysis of real-life examples.

Assessing Exposure-Time Treatment Effect Heterogeneity in Stepped Wedge Cluster Randomized Trials

Rui Wang
Harvard Medical School
rwang@hsph.harvard.edu
Coauthors: Lara Maleyeff, Harvard School of Public Health; Fan Li, Yale University; Sebastien Haneuse, Harvard School of Public Health

A stepped wedge cluster randomized trial is a unidirectional crossover study in which timings of treatment initiation for clusters are randomized. Because the timing of treatment initiation is different for each cluster, an emerging question is whether the treatment effect depends on the exposure time, namely, the time duration since the initiation of treatment. Existing approaches for assessing exposure-time treatment effect heterogeneity either assume a parametric functional form of exposure time or model the exposure time as a categorical variable, in which case the number of parameters increases with the number of exposure-time periods, leading to a potential loss in efficiency. In this article, we propose a new model formulation for assessing treatment effect heterogeneity over exposure time. Rather than a categorical term for each level of exposure time, the proposed model includes a random effect to represent varying treatment effects by exposure time. This allows for pooling information across exposure-time periods and may result in more precise overall and exposure-time specific treatment effect estimates. In addition, we develop an accompanying permutation test for the variance component of the heterogeneous treatment effect parameters. We conduct simulation studies to compare the proposed model and permutation test to alternative methods to elucidate their finite-sample operating characteristics, and to generate practical guidance on model choices for assessing exposure-time treatment effect heterogeneity in stepped wedge cluster randomized trials.

Topological Inference on Brain Signals

Yuan Wang
University of South Carolina
wang578@mailbox.sc.edu

Statistical inference of brain signals from diverse clinical groups often requires considerable technicality and computational power. Incorporating in statistical inference the shape of brain signals decoded with persistent homology allows us to tackle the heterogeneity of the data more effectively. This talk showcases recent methodological development and empirical findings on topological inference of brain signals from various experiments.

Bayesian Pooled Testing Regression With Measurement Error

Md Shamim Sarker
Radford University
msarker@radford.edu
Coauthors: Joshua Tebbs, University of South Carolina; Laura Hungerford, Virginia Tech

Pooled testing is widely used for screening and surveillance of diseases of low prevalence. While such testing is often valued for efficiency in identifying cases, regression approaches, which relate pooled responses to individual covariates can yield population-level estimates of disease positivity. Although statistical methods for these estimates have been widely studied, most work has developed under the assumption that covariates are measured without error, which is rarely true in real world applications. We describe new regression methods that provide reliable estimates and inference from pooled testing data with potentially mismeasured covariates. We consider a general modeling framework to accommodate test outcomes from any pooled testing protocol, where measurement error is corrected through a latent structural model. We also develop a hypothesis test to identify any violation of the latent model specification. This approach provides improved estimates of the prevalence from pooled samples for public health decision-making.

Modeling Negatively Skewed Survival Data in Accelerated Failure Time Models
Sophia Waymyers
Francis Marion University
swaymyers@fmarion.edu
Coauthor: Hrishikesh Chakraborty, Duke University

Negatively skewed survival data occasionally arises in medical research. We examine the efficacy of the reflected-shifted-truncated gamma (RSTG) distribution in accelerated failure time (AFT) models with and without individual frailty using maximum likelihood methods and an expectation-maximization algorithm. Using simulated negatively skewed data and information theoretic criteria, we show that the RSTG AFT model performs better than the AFT model with baseline exponential, generalized F, generalized gamma, Gompertz, log-logistic, lognormal, Rayleigh or Weibull distribution. While our methods are motivated by and applied to pediatric nephrotic syndrome data, they apply to other studies that collect negatively skewed data.

Hidden Population Estimation With Auxiliary Information
Justin Weltz
Duke University
justin.weltz@duke.edu
Coauthors: Alexander Volfovsky, Eric Laber

Many populations defined by illegal or stigmatized behavior are difficult to sample using conventional survey methodology. Respondent Driven Sampling (RDS) is a participant referral process frequently employed in this context to collect information without violating privacy. This sampling methodology operates through a stochastic process on a social network and generates a partially observed graph. Previous methods have attempted to estimate missing edges using participant arrival times. Unfortunately, these subgraph estimates exhibit highly biased edge set sizes, causing problems for downstream inference. We introduce a new procedure motivated by concepts from Indirect Inference to estimate the hidden population size using a more accurate complete subgraph. We also propose collecting additional network information during RDS to improve this procedure. These advances result in better estimation of the hidden population size over a range of simulation contexts.

Signal-To-Noise Ratio Aware Minimaxity for Sparse Gaussian Sequence Models
Haolei Weng
Department of Statistics and Probability, Michigan State University
wenghaol@msu.edu
Coauthors: Yilin Guo and Arian Maleki, Department of Statistics, Columbia University

Estimating a sparse vector under Gaussian sequence model is a fundamental problem in nonparametric and high-dimensional statistics. Classical results showed that both hard-thresholding and soft-thresholding estimators are asymptotically minimax optimal. However, such minimax optimality is inadequate to explain different (and often sub-optimal) performances of the aforementioned estimators in various signal-to-noise ratio (SNR) settings, as it focuses on the most challenging part of the parameter space. To address this issue, we consider a refined minimax framework where factors such as SNR and sparsity level influencing the hardness of the problem are carefully monitored. By a delicate second-order asymptotic analysis, we reveal three regimes in which distinct estimators achieve minimax optimality. In particular, hard-thresholding estimator outperforms soft-thresholding estimator and remains (asymptotically) minimax optimal in the high SNR regime; as SNR decreases, new optimal estimators will emerge. These new theoretical findings are much more informative towards understanding the sparse estimation problem in practice.

Maximum One-Factor-At-A-Time Designs for Screening in Computer Experiments
Qian Xiao
University of Georgia
qx69137@uga.edu
Coauthors: V. Roshan Joseph, Georgia Institute of Technology; Douglas M. Ray, US Army-CCDC Armaments Center

Identifying important factors from a large number of potentially important factors of a highly nonlinear and computationally expensive black box model is a difficult problem. Morris screening and Sobol' design are two commonly used model-free methods for doing this. In this article, we establish a connection between these two seemingly different methods in terms of their underlying experimental design structure and further exploit this connection to develop an improved design for screening called Maximum One-Factor-At-A-Time (MOFAT) design. We also develop a fast construction algorithm for MOFAT design using a novel transformation method. Several examples are presented to demonstrate the advantages of the proposed design compared to Morris screening and Sobol' design methods.

Constructing Covariance Functions for Axially Symmetric Processes on the Sphere
Xiaohuan (Max) Xue
UNC Greensboro
x_xue2@uncg.edu

Covariance functions are used to characterize dependency in spatial statistics and the construction of covariance functions is critical when performing prediction or "kriging". In this talk, we will discuss the the construction of covariance models for axially symmetric processes on the sphere. We will first review some of the recent development in this area, and we propose our preliminary results, then conclude the presentation with a discussion of future work.

Selecting Nearly Optimal Subdata
Min Yang
University of Illinois at Chicago
minyang.stat@gmail.com
Coauthor: Yike Tang

Big data brings unprecedented challenge of analyzing such data due to its extraordinary size. One strategy of analyzing such massive data is data reduction. Instead of analyzing the full dataset, a selected subdata set is analyzed. Various subdata selection methods have been proposed. Those methods are often based on the characterization of an optimal design. There are significant differences between optimal design and subdata selection: (i) an optimal design point may not exist in a given full data and (ii) while a point can be selected multiple times in an optimal design, it can only be selected once in a subdata selection. While the trade-off between computation complexity and statistical efficiency has been studied, little is known how efficient the selected subdata is in terms of statistical efficiency. To answer this question, we need to find an optimal subdata. Deriving an optimal subdata, however, is a N-P hard problem. In this talk, a novel framework to derive a nearly-optimal subdata, under any given statistical model, regardless of optimality criterion or parameters of interest, will be introduced. This framework has three benefits: (i) it shows us the structure of a nearly-optimal subdata for any given full data under various set-ups (model, optimality criterion, parameter of interest); (ii) it measures highly accurate statistical efficiency; and (iii) it provides a tool of deriving a nearly optimal subset in active learning where statistical efficiency is the main concern.

CW_ICA: An Efficient Dimensionality Selection Method for Independent Component Analysis

Yuyan Yi
Department of Mathematics and Statistics, Auburn University
yzy0080@aubrun.edu
Coauthors: Jingyi Zheng, Nedret Billor

Independent component analysis (ICA) is one of the most commonly used blind source separation (BSS) techniques for signal preprocessing, such as noise reduction and feature extraction. The main parameter in the ICA method is the number of independent components (IC) that is a crucial step before any modeling. Extracting too few ICs can lead to impure estimated signals, which may still contain mixed signals (under-decomposition), whereas choosing a large number of ICs might cause signal-to-noise deterioration and overfitting of the source signal (over-decomposition).

Although several methods have been proposed for the determination of the number of ICs, they have challenges, such as robustness and instability, which have not been studied or addressed in the literature. Robustness is referred to whether the presence of contaminated or anomalous observations in a dataset would influence the outcome of the method, which is the optimal number of ICs in our case. Moreover, depending on the data characteristics, different ICA methods, e.g. fastICA, Infomax, etc., are used by the researchers. However, existing methods for determining the number of ICs are not applicable to all ICA methods, which introduces additional challenges (e.g., uncertainty and instability) to choose the optimal number of ICs.

Therefore, to address the aforementioned issues, we propose a novel method, named the column-wise independent component analysis (CW_ICA), to determine the optimal number of ICs. The main idea behind CW_ICA is to measure the relationship between ICs from two different blocks, which are generated by randomly splitted mixture signals. The quantitative measurement of the ICA model with q ICs is defined as $R_q = \min_{1\le i\le q} \max |r_i|$, which describes the smallest column-wise maximum value in an off-diagonal correlation matrix, and the r_i denotes the rank-based correlation coefficient, i.e., Spearman's ρ, which is robust.

With simulation and raw scalp electroencephalogram (EEG) signal data as a validation set, we compare the proposed CW_ICA with several existing methods through different ICA methods. Results show that the proposed CW_ICA is a reliable and stable method for determining the optimal number of components in ICA. This method is robust (using rank-based correlation coefficient), has a wide range of applications (i.e., EEG, liquid chromatography–mass spectrometry (LC–MS), etc.), and can be used in conjunction with a variety of ICA methods (i.e., fastICA, Infomax, etc.).

Query-Augmented Active Metric Learning

Yubai Yuan
Pennsylvania State University
yvy5509@psu.edu
Coauthors: Yujia Deng, Meta; Haoda Fu, Eli Lilly; Annie Qu, University of California, Irvine

In this talk we propose an active metric learning method for clustering with pairwise constraints. The proposed method actively queries the label of informative instance pairs, while estimating underlying metrics by incorporating unlabeled instance pairs, which leads to a more accurate and efficient clustering process. In particular, we augment the queried constraints by generating more pairwise labels to provide additional information in learning a metric to enhance clustering performance. Furthermore, we increase the robustness of metric learning by updating the learned metric sequentially and penalizing the irrelevant features adaptively. In addition, we propose a novel active query strategy that evaluates the information gain of instance pairs more accurately by incorporating the neighborhood structure, which improves clustering efficiency without extra labeling cost. In theory, we provide a tighter error bound of the proposed metric learning method utilizing augmented queries compared with methods using existing constraints only. We also investigate the improvement using the active query strategy instead of random selection. Numerical studies on simulation settings and real datasets indicate that the proposed method is especially advantageous when the signal-to-noise ratio between significant features and irrelevant features is low.

Statistical Methods for Interval-Censored Multistate Data and Mis-Measured Covariates With Application in HIV Care

Hongbin Zhang
University of Kentucky
hongbin.zhang@uky.edu

In 2015, WHO announced the Treat All policy which recommends immediate antiretroviral therapy (ART) treatment of HIV infected people, regardless of disease severity. In evaluation the impact of adopting theTreat All policy at national level, the relationship between the biomarkers such as CD4 counts and WHO clinical stages (1: asymptomatic; 2 mild; 3: advanced; 4: severe; 5: mortality) is investigated. The WHO clinical stage data are interval-censored as the exact time of stage to stage transition between the clinical visits is unobservable. The CD4 covariate can have substantial measurement error. We proposed statistical methods for event history data subject to interval-censoring and mis-measured time-varying covariates: 1) two-steps method where the prediction of the true time-varying covariates was plugged into the outcome model for the estimation; and 2) joint model methods in which parameters from the longitudinal covriates model and from the survival model were simultaneously estimated. The methods were applied to real-world service deliver data in Central Africa and evaluated with simulation.

An Optional Quantitative Mixture RRT Model that Accounts for Lack of Trust

Joia Zhang
University of Washignton, Seattle
joiaz@uw.edu
Coauthor: Sat Gupta, University of North Carolina Greensboro

In this study, we introduce an optional quantitative RRT model that combines the elements of both the Pollock and Bek (1976) additive RRT model and the Greenberg et al. (1971) unrelated question quantitative RRT model. We examine the utility of the proposed mixture model using a unified measure of efficiency and privacy introduced by Gupta et al. (2018). We also account for the lack of trust in RRT models. The results show that the mixture model outperforms the two component models.

High-Dimensional Spatial Quantile Function-on-Scalar Regression

Zhengwu Zhang
UNC Chapel Hill
zz10c@email.unc.edu
Coauthors: Xiao Wang, Purdue University; Linglong Kong, University of Alberta; Hongtu Zhu, UNC Chapel Hill

We introduce a novel spatial quantile function-on-scalar regression model, which studies the conditional spatial distribution of a high-dimensional functional response given scalar predictors. With the strength of both quantile regression and copula modeling, we are able to explicitly characterize the conditional distribution of the functional or image response on the whole spatial domain. Our method provides a comprehensive understanding of the effect of scalar covariates on functional responses across different quantile levels and also gives a practical way to generate new images for given covariate values. Theoretically, we establish the minimax rates of convergence for estimating coefficient functions under both fixed and random designs. We further develop an efficient primal-dual algorithm to handle high-dimensional image data. Simulations and real data analysis are conducted to examine the finite-sample performance.

PERCEPT: a New Online Change-Point Detection Method Using Topological Data Analysis
Xiaojun Zheng
Duke University
xz264@duke.edu
Coauthors: Simon Mak, Duke University; Liyan Xie, The Chinese University of Hong Kong, Shenzhen; Yao Xie, Georgia Institute of Technology

Topological data analysis (TDA) provides a set of data analysis tools for extracting embedded topological structures from complex high-dimensional datasets. In recent years, TDA has been a rapidly growing field which has found success in a wide range of applications, including signal processing, neuroscience and network analysis. In these applications, the online detection of changes is of crucial importance, but this can be highly challenging since such changes often occur in low-dimensional embeddings within high-dimensional data streams. We thus propose a new method, called PERsistence diagram-based ChangE-PoinT detection (PERCEPT), which leverages the learned topological structure from TDA to sequentially detect changes. PERCEPT follows two key steps: it first learns the embedded topology as a point cloud via persistence diagrams, then applies a non-parametric monitoring approach for detecting changes in the resulting point cloud distributions. This yields a non-parametric, topology-aware framework which can efficiently detect online geometric changes. We investigate the effectiveness of PERCEPT over existing methods in a suite of numerical experiments where the data streams have an embedded topological structure. We then demonstrate the usefulness of PERCEPT in two applications on solar flare monitoring and human gesture detection.

Targeted Maximum Likelihood Estimation With Network-Dependent Data
Paul Zivich
UNC Chapel Hill
zivich.5@gmail.com

Suppose the parameter of interest is the mean of an outcome under an investigator-specified distribution of actions, referred to as a policy, where the distribution of actions can be set as fixed values or probabilistically. To estimate this parameter, the assumption of no interference (an individual's potential outcome is independent of all other individuals' actions) is a potential hurdle, most clearly exemplified in infectious disease research. While less often discussed, concerns of interference and network dependence extend to a variety of other substantive areas, including cancer research, substance abuse, and education. Recently, targeted maximum likelihood estimation (TMLE) has been extended for setting with network-dependent data, referred to as network-TMLE. I will briefly review TMLE for independent data for the mean under a policy. Then I will detail how TMLE is extended for interference with a known dependence structure. Results for an extensive simulation study of network-TMLE under varying data generating mechanisms is then presented. Next, an application of network-TMLE to estimation of the influenza infection risk under alternative policies for influenza vaccination uptake among university students is presented. Finally, I conclude with remaining challenges to application of network-TMLE and open areas of research.