Quality and Productivity Research Conference
July 26 – July 29, 2021
Turnbull Conference Center
Tallahassee, FL

Data Science and Statistics for Quality

Table of Contents

Quality and Productivity Research Conference

An annual conference sponsored by the ASA Section on Quality and Productivity

Conference Mission

The mission of the Quality and Productivity Research Conference (QPRC) is to improve the quality of products and services and the productivity of industries by stimulating the research and development of better statistical methods for quality and productivity improvement. The conference emphasizes research and development of statistical methods driven by important applications and looks to identify new application areas where statistics can have a significant impact. The conference includes statistical topics such as case studies as well as non-statistical topics such as writing, communication, and management that have an important influence on the effectiveness of statistical applications.

Conference Background

In 1984, AT&T Bell Laboratories organized the first Quality and Productivity Research (Mohonk) Conference. This first conference, and others that followed, stimulated many statisticians to develop improved methods for product and process design, manufacturing, and other business processes. Typical conference participants include industrial and academic statisticians, quality professionals, engineers, statistics and engineering students, and technical managers.

Scholarships

Student participation in the QPRC is strongly encouraged. Two Mary G. and Joseph Natrella Scholarships are awarded at the conference each year and include funds for participation in the conference in addition to the general scholarship funds. The scholarships are funded by the Q&P Mary G. and Joseph Natrella Scholarship Fund and the Quality and Productivity Research Conference.

In addition, several QPRC Student Scholarships are awarded each year to support student participation in the conference. Some of these scholarships generally are funded by the Quality and Productivity Research Conference. Since 2017 the QPRC also has been awarded annual NSF Grants which provide significant additional student support for both the Quality and Productivity Research Conference and the Conference Short Course.

Conference Honoree and Plenary Speakers

Geoff Vining
Conference Honoree

This year's conference honoree is Dr. Geoff Vining. Geoff Vining is a Professor of Statistics at Virginia Tech, where from 1999 – 2006, he also was the department head. He holds an Honorary Doctor of Technology from Luleå University of Technology. He is an Honorary Member of the ASQ (the highest lifetime achievement award in the field of Quality), an Academician of the International Academy for Quality, a Fellow of the American Statistical Association (ASA), and an Elected Member of the International Statistical Institute. He is the Founding Chair of the International Statistical Engineering Association (ISEA). He is a founding member of the US DoD Science of Test Research Consortium.

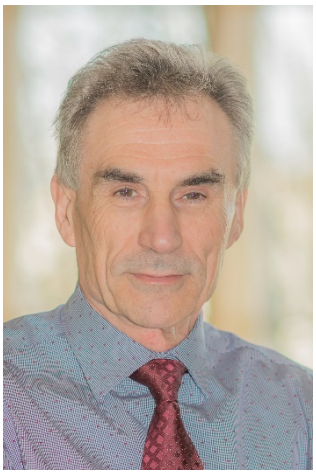
Dr. Vining won the 2010 Shewhart Medal, the ASQ career award given to the person who has demonstrated the most outstanding technical leadership in the field of modern quality control. He also received the 2015 Box Medal from the European Network for Business and Industrial Statistics (ENBIS). This medal recognizes a statistician who has remarkably contributed to the development and the application of statistical methods in European business and industry. In 2013, he received an Engineering Excellence Award from the NASA Engineering and Safety Center. He received the 2011 William G. Hunter Award from the ASQ Statistics Division for excellence in statistics as a communicator, consultant, educator, innovator, and integrator of statistics with other disciplines and an implementer who obtains meaningful results.

Dr. Vining is the author of three textbooks. He is an internationally recognized expert in the use of experimental design for quality, productivity, and reliability improvement and in the application of statistical process control. He has extensive consulting experience, most recently with the U.S. Department of Defense through the Science of Test Research Consortium and with NASA.

Peihua Qiu
Plenary Speaker

Peihua Qiu is a professor and the founding chair of the Department of Biostatistics at the University of Florida. He received his Ph.D. in statistics from the Department of Statistics at the University of Wisconsin at Madison in 1996. He then worked as a senior research consulting statistician for the Biostatistics Center at the Ohio State University during 1996-1998, and as a faculty member of the School of Statistics at the University of Minnesota during 1998-2013. Qiu has made substantial contributions in the research areas of jump regression analysis, image processing, statistical process control, survival analysis, and disease screening and surveillance. So far, he has published two research monographs and over 120 research papers in refereed journals. He is an elected fellow of the American Statistical Association, an elected fellow of the Institute of Mathematical Statistics, and an elected member of the International Statistical Institute. He served as associate editor for a number of top statistical journals, including JASA, Biometrics, and Technometrics. He was the editor of Technometrics during 2014-2016.

William Welch
Plenary Speaker



Will Welch joined the Department of Statistics, UBC as a Professor in 2003, and was Head of Department from 2003 until 2008 and Interim Head August-December 2018. Prior to that he was at the University of Waterloo for 16 years. He also holds the honorary title of Visiting Professor in the Business School, Loughborough University, UK.

Welch's research spans computer-aided design of experiments, quality improvement, the design and analysis of computer experiments, statistical methods for drug discovery, and machine/statistical learning. Please see this link for publications.

Welch has served on the editorial boards of the Annals of Applied Statistics, the Canadian Journal of Statistics, and the SIAM/ASA Journal on Uncertainty Quantification. He has also served as President of the Business and Industrial Statistics Section of the Statistical Society of Canada and as Associate Director of the Canadian Statistical Sciences Institute (CANSSI). He has won the American Statistical Association's Statistics in Chemistry Prize and is a Fellow of the American Statistical Association.

Statistical Thinking for Industrial Problem Solving

A free online statistics course

In virtually every field, deriving insights from data is central to problem solving, innovation and growth. But without the right skills, those insights remain out of reach. That's why we created this free online course. Learn core statistical methods, when to apply them, and how to interpret and communicate results.

jmp.com/statisticalthinking

jmp®
Statistical Discovery.™ From SAS.

Seven self-paced modules include instructional videos, JMP® demonstrations, questions and practices.

Statistical Thinking and Problem Solving
Learn how to map a process, define and scope your project, and determine the data you need to solve your problem.

Exploratory Data Analysis
Learn how to describe data with graphics and use interactive visualizations to find and communicate the story in your data.

Quality Methods
Learn about tools to quantify, control and reduce variation in your product, service or process.

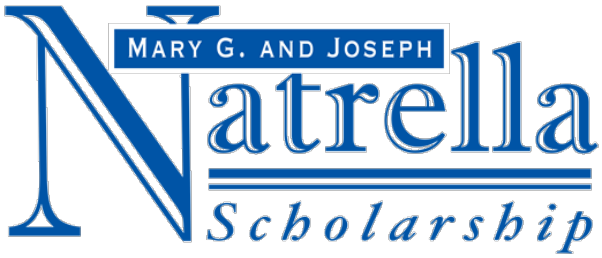
Decision Making With Data
Learn to draw inferences from data, construct statistical intervals, perform hypothesis tests, and understand the relationship between sample size and power.

Correlation and Regression
Learn how to study the linear association between pairs of variables, and how to fit and interpret linear and logistic regression models.

Design of Experiments
Learn the language of design of experiments (DOE) and see how to design, conduct and analyze an experiment in JMP.

Predictive Modeling and Text Mining
Learn how to identify possible relationships, build predictive models and derive value from free-form text.

Natrella Scholarship



This scholarship was initiated by a contribution to the ASA Quality and Productivity Section given by Joseph Natrella at the time of Mary Natrella's death to honor her many contributions to the statistical community. The Natrella's always maintained a strong mutual interest in quality applications of statistics. Mary was for many years a staff member of the Statistical Engineering Division of the National Institute of Standards and Technology (NIST). Joe's career was primarily with the Department of Defense and NASA as a mathematician in charge of data processing and computations.

Mary's most important publication, NBS Handbook 91: Experimental Statistics, is one of the all-time best selling publications of NIST. Originally published in 1963 by the Government Printing Office, it was later reprinted by Wiley-InterScience in its Selected Government Publications series, and also has been reprinted in paperback by Dover Publications. The ASA Q&P Section established the scholarship in 2000 to honor Mary's 36 years as an author, teacher and consulting statistician.

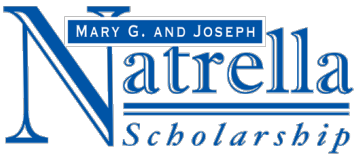
Natrella Scholarship Winners – 2021

Siddhesh Kulkarni is PhD Candidate in Biostatistics at University of Louisville. He earned his MS Statistics from University of Connecticut in 2018. His thesis is based on the theme of developing Bayesian Methodologies for constrained spaces. His current research focuses in the areas of high dimensional covariance estimation and graphical models, theoretical properties of Bayesian Computation/MCMC, stochastic processes etc. Along with academics he is quite active in various leadership roles in professional as well as social/community services. He is a recipient of multiple awards for his academic achievements including this years' ASA Biopharmaceutical Sections' Scholarship and the Natrella Scholarship.

Minhee Kim received the B.S. degree in industrial and management engineering from Pohang University of Science and Technology (POSTECH), Pohang, Korea, in 2017 and the M.S. degree in statistics from the University of Wisconsin-Madison, Madison, WI, USA, in 2021. Currently, she is working towards the Ph.D. degree at the Department of Industrial and Systems Engineering, University of Wisconsin-Madison. Her research interests are system automation, quality control, and predictive analysis with applications in manufacturing and healthcare.

Conference Sponsors

Many thanks to our generous sponsors!



Florida State University Support

We are pleased to acknowledge the outstanding support from several units within Florida State University:

- Laurel Fulkerson, Vice President for Research
- Sam Huckaba, Dean of the College of Arts and Sciences
- Murray Gibson, Dean of the College of Engineering
- Xufeng Niu, Chair of the Department of Statistics
- Okenwa Okoli, Chair of the Department of Industrial and Manufacturing Engineering

Short Course

Visualize, Wrangle, Present: Work Effectively with the Tidyverse and RMarkdown

The conference short course will be held in-person and remotely on July 26.

Course Description:

The Tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

In this workshop, we will introduce you to basic Tidyverse concepts and help you work more effectively with your analyses.

Prerequisites: Basic familiarity with R is suggested. We highly suggest going through the two RStudio Primers here before the workshop:

https://rstudio.cloud/learn/primers/1

Part 1: Visualization using ggplot2 and the grammar of graphics

ggplot2 is an extremely powerful visualization library. In this session, we will cover basic plots (barplots, scatterplots, and boxplots) and the ways to customize these visualizations using color.

Part 2: Summarizing, transforming, and cleaning data using dplyr

The tidyverse includes many functions that allow us to quickly group and summarize data to produce useful tabular summaries of a large dataset. We will cover useful functions in dplyr for summarizing data and working with missing values in a data. Our focus will be on utilizing streams of continuously updated data and ways to productively work with them.

Part 3: Reproducible Reporting with RMarkdown and Friends

RMarkdown is an extremely powerful file format that allows you to mix explanatory text, code, and figures/tables. From a single Rmarkdown file you can produce webpages, presentations, and even nicely formatted emails. In this session, we will introduce RMarkdown syntax, how to output different formats, and resources for shaping and styling your outputs.

***Course Instructor*: Ted Laderas**

I am an Assistant Professor in the Division of Bioinformatics and Computational Biology in the Department of Medical Informatics and Clinical Epidemiology at OHSU (BCB/DMICE) and a member of the OHSU Knight Cancer Institute. Collaboration, training and team science are passions of mine. I believe that research should not be a lonely practice, and building communities

of practice in science and research that are psychologically safe and inclusive are the key to doing better, more robust science.

My free Ready for R course currently is at over 1300 external users, and I have built a learning community around it. The course has been lauded by users as being more inclusive and less intimidating than other R courses.

***Course assistant*: Eric Leung**

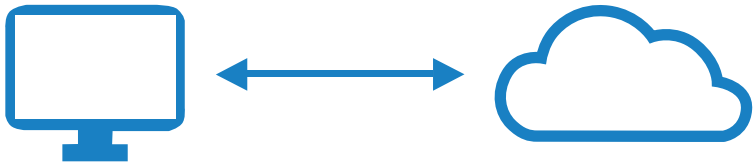
Minitab®

Powerful statistical software everyone can use

Minitab®

Access Minitab from Anywhere

Business doesn't stop, so why should your data analysis? Increase your productivity and collaboration with Minitab accessible through both your desktop and on the cloud with our web app.

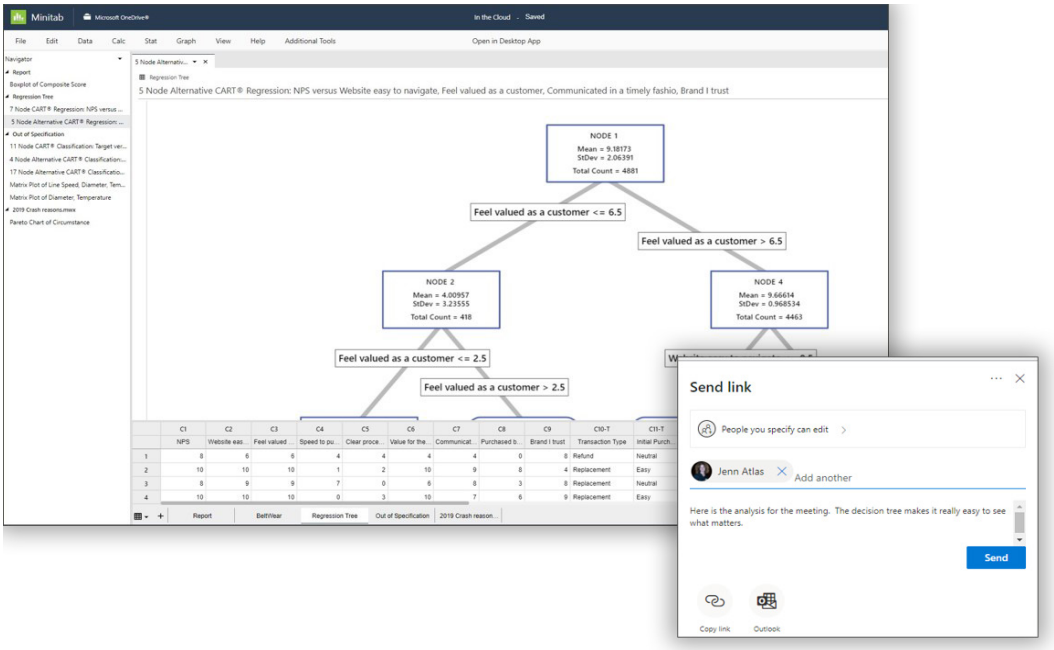


Log in from nearly any device to complete analyses, find insights and make better decisions whether you're:

in the office

at home

halfway around the world



Schedule

	QPRC 2021 Schedule			
	Monday	**Tuesday**	**Wednesday**	**Thursday**
	July 26	**July 27**	**July 28**	**July 29**
8:00 AM	Breakfast	Breakfast	Breakfast	Breakfast
8:30 AM				
9:00 AM	Short course	Plenary: Peihua Qiu	Session 4	Plenary: Geoff Vining
9:30 AM				
10:00 AM		Break	Break	Session 7
10:30 AM		Session 1	Plenary: William Welch	
11:00 AM				
11:30 AM	Lunch		Demo	Break
12:00 PM		Lunch	Lunch	Session 8
12:30 PM	Short course			
1:00 PM		Session 2	Session 5	
1:30 PM				Closing Remarks and Lunch
2:00 PM				
2:30 PM	Break	Break	Break	
3:00 PM	Short course	Session 3	Demo	
3:30 PM			Session 6	
4:00 PM				
4:30 PM		Posters	Tour and Reception: NHMFL / HPMI	
5:00 PM				
5:30 PM				
6:00 PM		Banquet		
6:30 PM				
7:00 PM				
7:30 PM				
8:00 PM				


NISS
National Institute of
Statistical Sciences


R
Studio®


TriloByte®
Ltd
STATISTICAL
SOFTWARE

Turnbull Conference Center Map

Conference Rooms

Monday (short course):
- 103 Dining
- 214 Classroom

Tuesday – Thursday:
- 108 Dining
- 103, 208, 214 Sessions

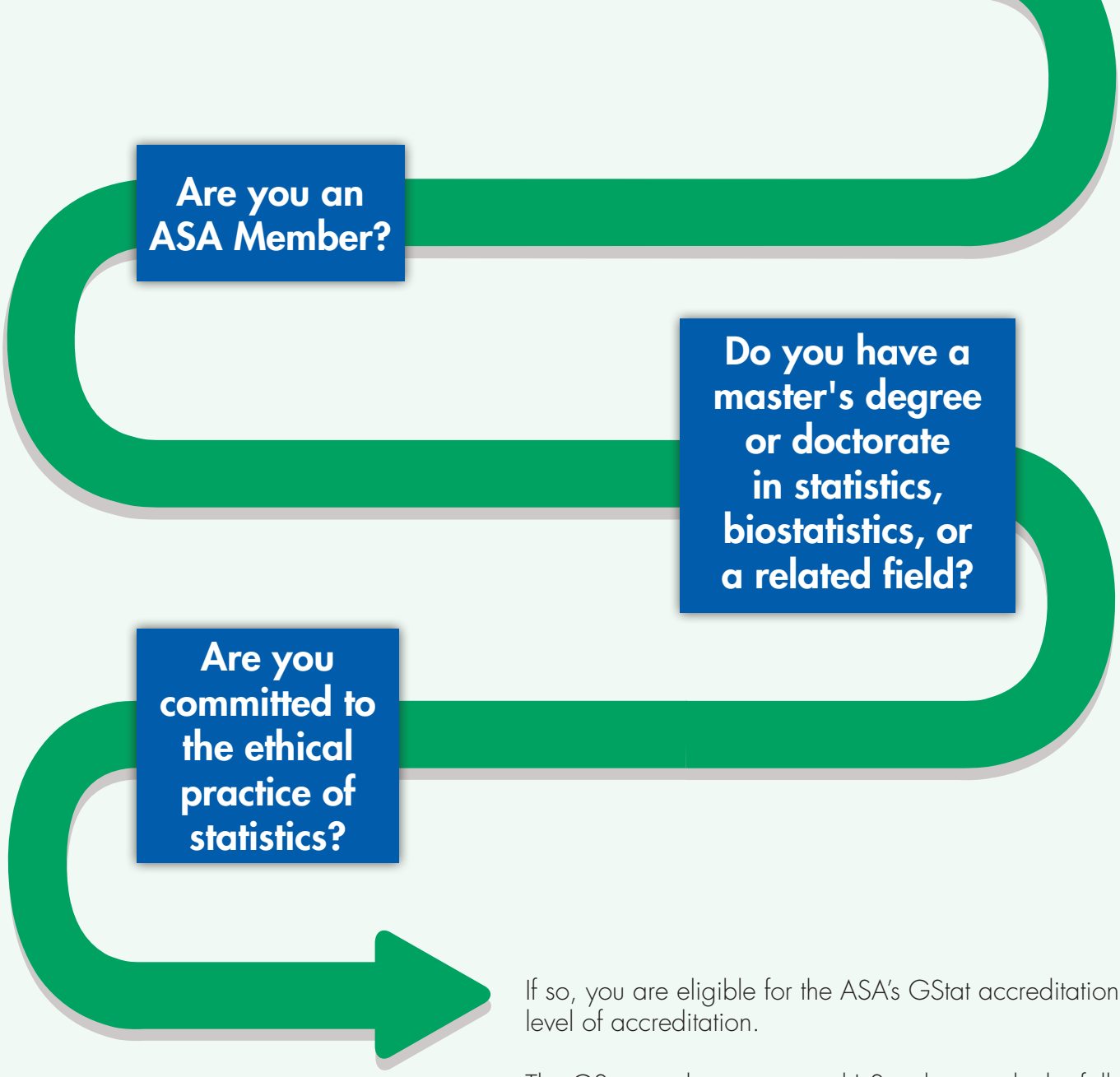
Organizing Committee

Local Committee (Florida State University):

Felipe Barrientos
Jon Bradley
Hongyuan Cao
Eric Chicken (Chair)
Chao Huang
Lifeng Lin
Qing Mai
Pamela McGhee
Chiwoo Park
Danice Stetson
Jonathan Stewart
Arda Vanli
Hui Wang

QPRC Steering Committee:

Haim Bar
Michael Baron
Eric Chicken
Anne Driscoll
Martha Gardner
Blan Godfrey
Will Guthrie
Jeff Hooper (Chair)
Dan Jeske
Jami Kovach
Scott Kowalski
Christina Mastrangelo
Alexandra Piryatinska
Matt Plumlee
Sharad Prabhu
Angela Schoergendorfer
Mia Stephens
Nathaniel Stevens
Paul Tobias
Brian Weaver
Joanne Wendelberger
Emmanuel Yashchin



If so, you are eligible for the ASA's GStat accreditation—an entry level of accreditation.

The GStat application is quick! Simply provide the following:

- Contact information
- Résumé or CV
- A list of degrees and courses
- Demographic information (optional)

GStat holders are granted access to the same benefits as PStat® holders:

- Free access to LearnStat on Demand (online professional development courses)
- Reduced registration fees for ASA conferences, including JSM and CSP
- Special recognition at JSM, including a ribbon designation and exclusive reception

Also, GStat holders can request the ASA Accreditation Committee to review their progress toward meeting PStat® requirements.

gstat
Graduate Statistician

www.amstat.org/accreditation

BIGTENT statistics

Poster Program, Tutorial, and Schedule

This year's poster program will be virtual using Gather (https://www.gather.town/). A short tutorial is provided on the following pages.

The posters will be judged this year with a $100 Amazon gift certificate to be awarded to the student with for the QPRC 2021 Best Poster.

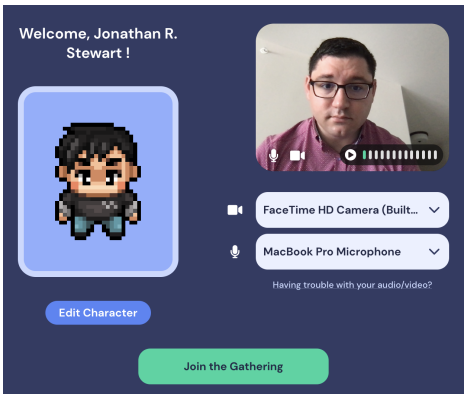
Many thanks to Dr. Jonathan Stewart for organizing the poster session.

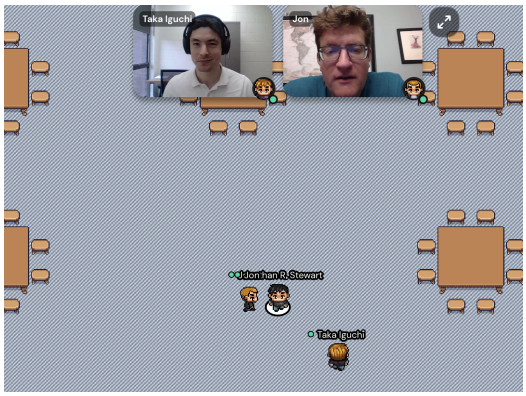
QPRC 2021
Gather Town tutorial for poster session

What is Gather Town?

Gather Town (https://gather.town/) is a 2-dimensional interactive video conferencing platform that can be accessed through either a web browser or a desktop application.

In Gather Town, you are able to choose an *avatar* that represents you. Once you have set-up your name, avatar, and video / audio preferences, you will be able to enter the virtual conference hall and participate in the poster session.





In the virtual conference hall, you will find posters, tables, and lounges, all with special functionalities to enable a successful and productive virtual poster session. The next sections of this tutorial will introduce you to some of the main functionality you will need to participate in the virtual poster session. When your avatar is close to other avatars, you can see and hear those persons' video and audio.

Accessing the virtual conference hall.

The easiest way to get up and running is to access Gather Town from your web browser. There is also a desktop application that can be downloaded, but it is recommended to simply access the platform via your browser. The virtual conference hall can be accessed by copy and pasting the following URL into your browser:

https://gather.town/app/jKOBKytGaES6ZONc/QPRC2021PosterSession

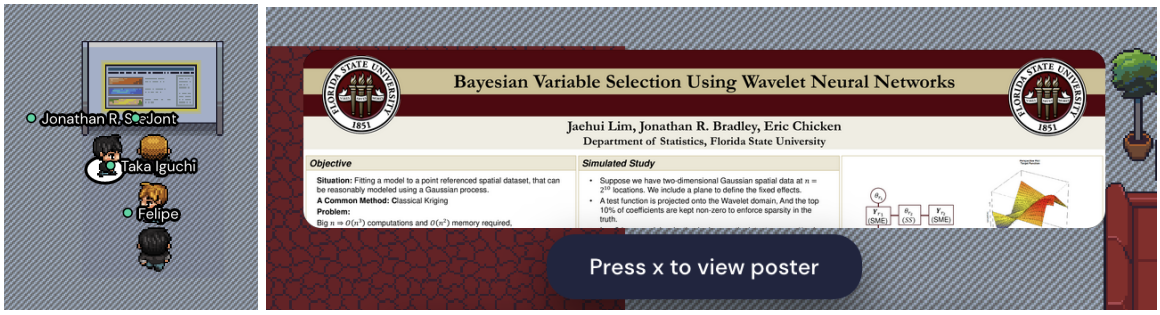
Password: QPRC2021

You will use this link and the password on the day of the poster session to access the virtual conference hall.

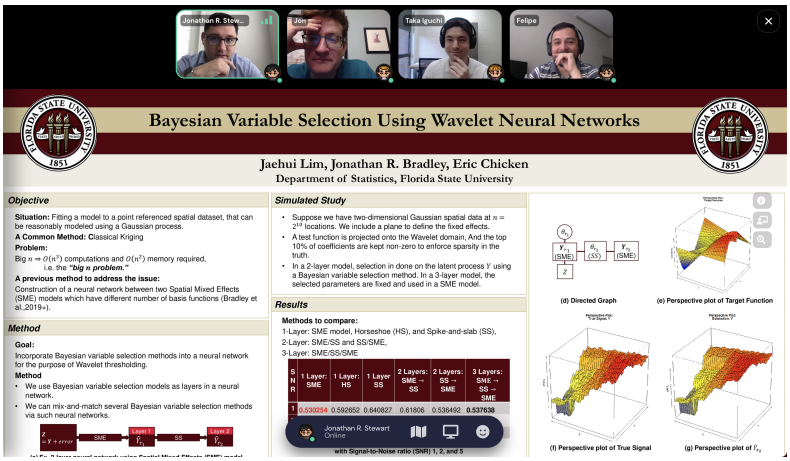
How to interact with Gather Town during the poster session.

The controls for Gather Town are relatively straightforward. You will use the arrow keys ↑↓←→ to control your avatar and have them walk around the conference and interact with posters and other conference participants.

Individuals posters are represented by a virtual image of a poster. The shadowy region in front of the poster is a *private space* in which you can chat with other participants also standing in front of the poster. When standing in a private space, you will only see video and hear audio from the other participants in that space. This allows you to communicate with the poster presenter unobstructed. When you are ready to go to another poster, simply use the arrow keys to walk your avatar to the next poster's region.



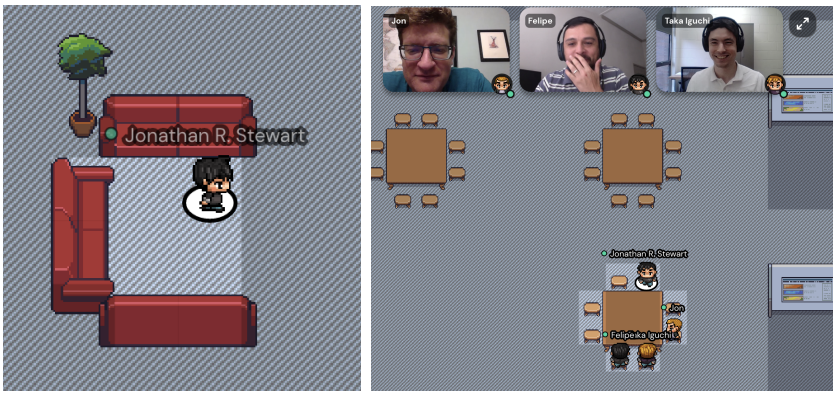
Once your avatar is standing in front of a poster in the shaded private space, you are able to press the x button on your keyboard to view the poster. A small preview of the poster will be visible which should tell you the poster title and the authors of the poster. This will bring up the poster, as well as the video chat boxes of all participants in the poster's private space.



While viewing a poster, you are able to use your mouse or trackpad to scroll up and down to view different parts of the poster, as well as zoom in and out. There are buttons arranged in a column on the right of the screen (highlighted above in a bright red box) while viewing a poster that let you get more information, use a pointer to create a virtual laser pointer effect, and buttons to zoom in and zoom out.

Using the private conversation spaces in the conference hall.

A number of private conversation spaces are setup in the virtual conference hall. These can be accessed just as the poster spaces were accessed. Simply use the arrow keys $\uparrow\downarrow\leftarrow\rightarrow$ on your keyboard to control your avatar and move them to one of the designated private spaces.



The private conversation spaces will appear as either a lounge (left) or a table (right). If any person's avatar is standing in the light shaded region, that person will be able to interact (video and audio) with any other persons whose avatars are in the private space. The private spaces are useful when you want to have a quiet conversation with one or more persons without interference from the conference hall, just as you would utilize these spaces in person.

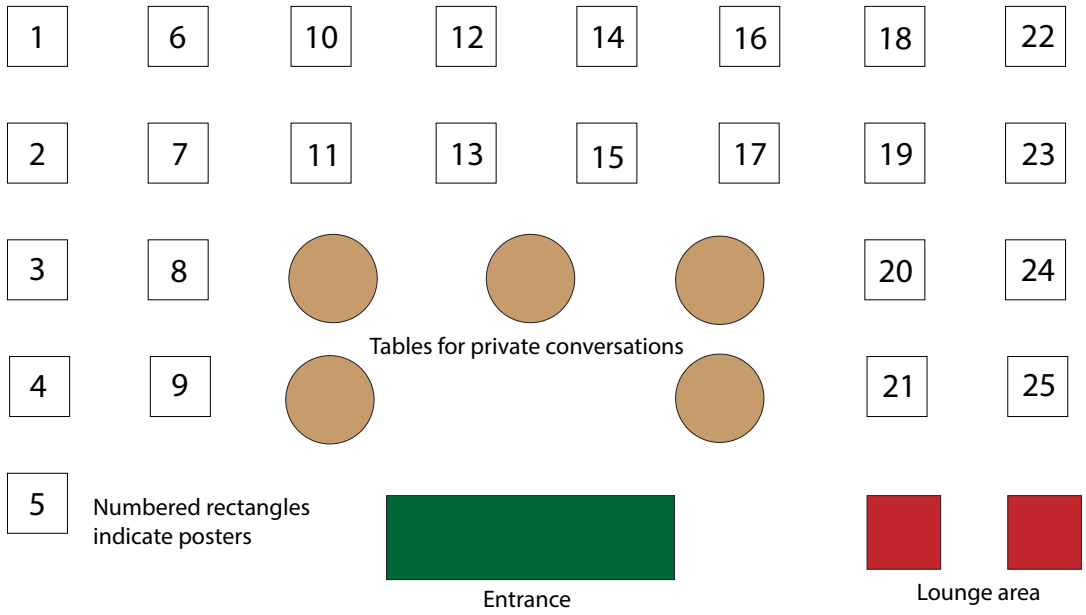
Try out the platform for yourself!

You may go to `https://gather.town/` to try out the Gather Town platform for yourself to get a hang of how to use it. The live walkthrough (click link "Try out live walkthrough" when visiting `https://gather.town/`) even has an example exhibition hall with a poster session if you follow the set path the developers have made.

We want the poster session to be as accessible and productive for all attendees as possible. If you have any questions about using the Gather Town platform or the poster session in general, please do not hesitate to email Jonathan Stewart at jrstewart@fsu.edu. We look forward to *seeing* you all there!

QPRC 2021
Poster Session Information

Virtual conference hall layout:



Poster session program:

1. Online sequential monitoring of spatio-temporal disease incidence rates
   Presented by Kai Yang

2. Tornado Property Loss Scale
   Presented by Thilini Mahanama

3. A generic online nonparametric monitoring and sampling strategy
   for high-dimensional heterogeneous processes
   Presented by Honghan Ye

4. NP-ODE: Neural process aided ordinary differential equations
   for uncertainty quantification of finite element analysis
   Presented by Yinan Wang

5. General penalized logistic regression for gene selection
   in high-dimensional microarray data classification
   Presented by Derrick Bonney

6. Neural network Gaussian process considering input uncertainty
   and its application to composite structures assembly
   Presented by Cheolhei Lee

7. Constrained minimum energy designs
   Presented by Chaofan Huang

8. Transparent sequential learning for statistical process control
   Presented by Xiulin Xie

9. A hybrid regression-ranking model with application in personalized radiomics
   Presented by Lingchao Mao

10. HODOR: A two-stage hold-out design for online randomized experiments
    Presented by Nicholas Larsen

11. A novel statistical data fusion framework for spatial predictive modeling with application in precision medicine
    Presented by Lujia Wang

12. Additive tensor decomposition considering structural data information
    Presented by Shancong Mou

13. Policies review of data science in industry and government: Case of SAARC countries
    Presented by Dila Bhandari

14. Digital twin development and maximum stress prediction for composite fuselage assembly
    Presented by Tim Lutz

15. Framework for a bayesian EWMA control chart using loss functions
    Presented by Chelsea Jones

16. Bayesian variable selection through wavelet neural network for spatio-temporal data
    Presented by Jaehui Lim

17. Nonparametric denoising of image sequences
    Presented by Fan Yi

18. A multivariate quality assurance approach for credit card customers and some features
    Presented by Mian Adnan

19. Structural tensor-on-tensor regression with interaction effects and its application to a hot rolling process
    Presented by Huihui Miao

20. Robust coupled tensor CP decomposition using ADMM method
    Presented by Meng Zhao

21. A sequence graph transform based method for monitoring discrete sequence processes
    Presented by Meserret Karaca

22. Understanding power grid network vulnerability through the stochastic lens of network motif evolution
    Yuzhou Chen

23. SAGE: Stealthy Attack GEneration for cyber-physical systems
    Presented by Michael Biehler

24. Building a risk adjusted model for monitoring psychopharmaceutical prescription for children in the child welfare system
    Presented by Julia Thome

25. TBD

Session Protocols

QPRC 2021 is a hybrid conference. Attendees will be participating both in-person and remotely via zoom. This presents some challenges for our presentation sessions. To promote orderly sessions, the organizing committee has come up with the procedures outline below.

In-person Attendees

In-person presenters will use a podium / computer station at the front of each room. For ease of loading, presentations should be in PDF format on a USB drive. The computer at the podium is a windows machine that is directly connected to the in-room display and zoom. There is a camera in the room that is focused on the speaker for the zoom audience.

In-person audience members will be able to see the speaker at the podium as well as on a large screen in the session room. A microphone will be in each room for questions. We ask that you use the microphone to pose questions to the speaker. This will ensure that remote participants will be able to hear your questions. When the speaker is remote, the in-room audience members will be able to see and hear the speaker on the room screen. Questions can be posed to remote speakers via chat (if you have a laptop with you) or by using the in-room microphone.

Remote Attendees

Remote presenters will connect to the appropriate zoom link before their talk. When it is their time to speak within a session the chair will give the speaker the ability to share their screen.

Remote audience members will be muted during the presentation. To ask a question, please post it in chat. Session chairs and in-room hosts will moderate questions for speakers.

Harassment-free Environment

Florida State University takes a strong stance against harassment and provides a harassment-free environment. This conference is on Florida State University property and will adhere to its official policies:

Sexual misconduct, including sexual harassment, sexual violence (rape, sexual assault, domestic violence, dating violence, & stalking), and all other forms of sex discrimination are violations of University policy and contrary to the University's values, which recognize the dignity and worth of each person. They are also illegal. Sexual misconduct will not be tolerated by Florida State University, whether by faculty, staff, students, visitors, or others.

The University expressly prohibits unlawful discrimination, harassment, or retaliation, whether in assumption, attitudes, acts, or policies. Conduct that intimidates by threat, brings about adversity, or creates a hostile environment, is contrary to the University's commitment of maintaining a harmonious, high performance work and educational environment.

More detailed descriptions are available at:

https://www.hr.fsu.edu/sections/equity-diversity-inclusion/sexual-misconduct-title-ix

and

https://policies.vpfa.fsu.edu/policies-and-procedures/faculty-staff/equal-opportunity-and-compliance-eoc#I3

Complaints may be made through these links:

https://www.hr.fsu.edu/sections/equity-diversity-inclusion/make-complaint

https://www.hr.fsu.edu/sites/g/files/upcbnu2186/files/DiscriminationComplaintForm_fill.pdf

Program

Tuesday, July 27

Session 1 A
Tuesday, July 27
10:30 a.m. – 12:00 p.m.
Room 103

Advancements in machine learning
Organizer and Chair: Qing Mai

Linear regression and its inference on noisy network-linked data
Tianxi Li, Department of Statistics, University of Virginia

A Screening Algorithm for Cross-Validated Kernel Support Vector Machines
Boxiang Wang, Department of Statistics and Actuarial Science, University of Iowa

Spectral clustering via adaptive layer aggregation for multi-layer networks
Haolei Weng, Department of Statistics and Probability, Michigan State University

Session 1 B
Tuesday, July 27
10:30 a.m. – 12:00 p.m.
Room 208

Some Statistical Monitoring Strategies for Modern-Day Processes
Organizer and Chair: Marcus Perry

A copula-based CUSUM scheme for monitoring auto-correlated processes
Yang-Li Liao, Department of Information Systems, Statistics and Management Science, University of Alabama

A distribution-free joint monitoring scheme for location and scale using individual observations
Marcus Perry, Department of Information Systems, Statistics and Management Science, University of Alabama

On the Simulation of Autocorrelated Processes
Zhi Wang, Bayer Crop Science

Detecting special-cause variation 'events' from process data signatures
Timothy M. Young, Department of Forestry, Wildlife and Fisheries, University of Tennessee

Session 1 C
Tuesday, July 27
10:30 a.m. – 12:00 p.m.
Room 214

Modeling and Testing
Organizer and Chair: Runze Li

A distribution free conditional independence test with application in causal discovery
Zhanrui Cai, Department of Statistics and Data Science, Carnegie Mellon University

Inference in High Dimensional Linear Measurement Error Models
Mengyan Li, Mathematical Sciences Department, Bentley University

Power-Enhanced Simultaneous Test of High-Dimensional Mean Vectors and Covariance Matrices
Xiufan Yu, Department of Applied and Computational Mathematics and Statistics, University of Notre Dame

Linear Hypothesis Testing in Linear Models with High Dimensional Responses
Runze Li, Department of Statistics, Pennsylvania State University

Session 2 A
Tuesday, July 27
1:00 p.m. – 2:30 p.m.
Room 103

Causal Inference with Survival Data
Organizer and Chair: Hongyuan Cao

The restricted mean time in favor of treatment
Lu Mao, Department of Biostatistics, University of Wisconsin-Madison.

Towards double robustness under the Cox marginal structure model
Ronghui (Lily) Xu, Department of Mathematics, University of California-San Diego

Estimation of causal quantile effects with a binary instrumental variable and censored data
Limin Peng, Department of Biostatistics, Emory University

A unified theory for covariate adjustment in the design and analysis of randomized clinical trials
Ting Ye, Department of Statistics, University of Pennsylvania

Session 2 B
Tuesday, July 27
1:00 p.m. – 2:30 p.m.
Room 208

Reliability of Big Data Integration
Organizer and Chair: Chao Huang

Multi-task Learning with High-Dimensional Noisy Images
Suprateek Kundu, Department of Biostatistics and Bioinformatics, Emory University

A Tree-based Federated Learning Approach for Personalized Treatment Effect Estimation from Heterogeneous Data Sources
Lu Tang, Department of Biostatistics, University of Pittsburgh

Debiasing Multidimensional Scaling
Qiang Sun, Department of Statistics, University of Toronto

Image-on-Scalar Partial Linear Single Index Model for Handling Heterogeneity in Imaging Studies
Chao Huang, Department of Statistics, Florida State University

Session 2 C
Tuesday, July 27
1:00 p.m. – 2:30 p.m.
Room 214

Statistical Process Control for Functional Data
Organizer and Chair: Takayuki Iguchi

Nonparametric multivariate profile monitoring using regression trees
Daniel Timme, Department of Statistics, Florida State University

Nonlinear profile monitoring with single index models
Takayuki Iguchi, Department of Statistics, Florida State University

Bayesian Sequential Monitoring of Density Estimates
Wright Shamp, Johnson and Johnson

Session 3 A
Tuesday, July 27
3:00 p.m. – 4:30 p.m.
Room 103

Change Point Detection Methodologies and Applications
Organizer and Chair: Zhe Wang

Efficient change detection in point processes driven by data-dependent random intensities
Moinak Bhaduri, Department of Mathematical Sciences, Bentley University

On the characteristics of the Generalized Shiryaev-Roberts procedure for the quickest change point detection in continuous time
Kexuan Li, Worcester Polytechnic Institute

Change Point Detection Methods for Renewable Energy Integration
Malaquias Pena, Department of Civil & Environmental Engineering, University of Connecticut

A Predictive Bayesian Approach to Sequential Time-Between-Events Monitoring
Sajid Ali, Department of Statistics, Quaid-i-Azam University

Session 3 B
Tuesday, July 27
3:00 p.m. – 4:30 p.m.
Room 208

Statistical Applications of Wavelets and Functional Data in the USAF
Organizer and Chair: Ray Hill

Experimental Design with a Functional Response
Matthew Scherer, Department of Operational Sciences, Air Force Institute of Technology

Wavelets on Graphs
Jeff Williams, Department of Operations Sciences, Air Force Institute of Technology

Wavelet Methods for Very-short term Forecasting of Functional Time Series
Jared Nystrom, Department of Operations Sciences, Air Force Institute of Technology

Session 3 C
Tuesday, July 27
3:00 p.m. – 4:30 p.m.
Room 214

Statistical Approaches and Machine Learning in Engineering Systems
Organizer and Chair: Hui Wang

Distortion Modeling and Compensation Across Materials and Processes in Laser-Based Additive Manufacturing Systems via Bayesian Neural Networks
Arman Sabbaghi, Department of Statistics, Purdue University

Sensor-based Modeling and Optimization of Additive Manufacturing
Hui Yang, Department of Industrial and Manufacturing Engineering, Pennsylvania State University

Concurrent Physics-Data-Driven Learning Methods for Resilient Cyber-Physical Systems
Olugbenga Moses Anubi, Department of Electrical & Computer Engineering, Florida A&M University

Multi-printer Co-learning of Kinematics-Induced Variations for Inter-connected Extrusion-based Additive Manufacturing
An-Tsun Wei and Hui Wang, Department of Industrial and Manufacturing Engineering, Florida State University

Wednesday, July 28

Session 4 A
Wednesday, July 28
9:00 a.m. – 10:00 a.m.
Room 208

Natrella Winners Presentations
Organizers and Co-Chairs: Will Guthrie and Scott Kowalski

Covariate Dependent Sparse Functional Data Analysis
Minhee Kim, Department of Industrial and Systems Engineering, University of Wisconsin-Madison

A Bayesian Approach for Joint Estimation for Sparse Canonical Correlation and Graphical Models
Siddhesh Kulkarni, University of Louisville

Session 4 B
Wednesday, July 28
9:00 a.m. – 10:00 a.m.
Room 214

A Brief Tutorial on the mixtools and tolerance Packages for R
Derek Young, Department of Statistics, University of Kentucky

Session 4 C
Wednesday, July 28
9:00 a.m. – 10:00 a.m.
Room 103

Sequentially monitoring nonlinear profiles using a Gaussian process model with heteroscedasticity
A. Valeria Quevedo, Department of Industrial and Systems Engineering, Universidad de Piura

Fitting time series models without identification: a hierarchical sparse proximal approach
Sam Davanloo, Department of Integrated Systems Engineering, Ohio State University

Demonstration Session 1
Wednesday, July 28
11:30 a.m. – 12:00 p.m.
Room 214

Minitab: Statistical Tools Everyone Can Use
Scott Kowalski and Jennifer Atlas, Minitab

Session 5 A
Wednesday, July 28
1:00 p.m. – 2:30 p.m.
Room 103

An Overview of Statistical Engineering
Organizer and Chair: Geoff Vining

NASA's Statistical Engineering Journey
Peter Parker, Sara Wilson, and Ray Rhew, National Aeronautics and Space Administration

Why are Statistical Engineers needed for Test & Evaluation? Lessons learned from my time spent in the desert, the air, windowless rooms, and sitting behind a computer.
Rebecca Medlin, Institute for Defense Analyses

Statistical Engineering and Wicked Problems?
Marcus Perry, Department of Information Systems, Statistics and Management Science, University of Alabama

Session 5 B
Wednesday, July 28
1:00 p.m. – 2:30 p.m.
Room 208

Uncertainty Quantification and Calibration of Computer Models
Organizer and Chair: Chiwoo Park

Using BART for Multiobjective Optimization of Noisy Multiple Objectives
Matthew Pratola, Department of Statistics, Ohio State University

Additive and projection pursuit models with Gaussian process regression
Rui Tuo, Department of Industrial and Systems Engineering, Texas A&M University

Bayesian MARS for efficient and informative nonlinear regression
Devin Francom, Statistical Sciences Group, Los Alamos National Lab

Inverse Problems in Materials Characterization
Sachin Shanbhag, Department of Scientific Computing, Florida State University

Session 5 C
Wednesday, July 28
1:00 p.m. – 2:30 p.m.
Room 214

Design and Analysis
Organizer and Chair:

Optimizing the Online User Experience: Design and Analysis of A/B Tests with Bayesian Comparative Probability Metrics
Nathaniel T. Stevens, Department of Statistics and Actuarial Science, University of Waterloo

Two-level orthogonal screening designs with 80, 96 and 112 runs: Construction and evaluation
Alan R. Vazquez, Department of Statistics, University of California, Los Angeles

Structure of Two-Level Nonregular Designs
David Edwards, Department of Statistical Sciences and Operations Research, Virginia Commonwealth University

Min-Max Optimal Design of Two-Armed Trials with Side Information
Qiong Zhang, School of Mathematical and Statistical Sciences, Clemson University

Demonstration Session 2
Wednesday, July 28
3:00 p.m. – 3:30 p.m.
Room 214

JMP Software: New Features in v16
Kevin Potcner, SAS Institute Inc.

Session 6 A
Wednesday, July 28
3:30 p.m. – 4:30 p.m.
Room 103

Univariate and multivariate techniques for root-cause analysis of quality
Henry Linder, Oden Technologies

Semi-Supervised Clustered Multi-Task Learning (SSC-MTL) for modeling connected systems
Sajjad Seyedsalehi, Department of Industrial & Operations Engineering, University of Michigan

Finding the source of grandma's chili: investigative text mining
Scott Lee Wise, SAS Institute, Inc.

Session 6 B
Wednesday, July 28
3:30 p.m. – 4:30 p.m.
Room 208

The road to scaling Hotelling multivariate control charts
Francois Dion, Dion Research LLC

On the performance of the generalized Shiryaev-Roberts control chart in continuous time
Aleksey S. Polunchenko, Department of Mathematical Sciences, Binghamton University

Bootstrapping quantiles of degradation data
Richard Warr, Department of Statistics, Brigham Young University

Session 6 C
Wednesday, July 28
3:30 p.m. – 4:30 p.m.
Room 214

Statistical methods applied to monitoring growth in shrimp farms
Ismael Sánchez, Department of Industrial and Systems Engineering, Universidad de Piura

An effective method for online disease risk monitoring
Lu You, Health Informatics Institute, University of South Florida

Monitoring processes with a large number of variables
Kevin Potcner, SAS Institute Inc.

Thursday, July 29

Session 7 A
Thursday, July 29
10:00 a.m. – 11:30 a.m.
Room 103

Defense Models and Testing
Organizer and Chair: Kelly Avery

Test Design Challenges in Defense Testing
Kelly Avery, Institute for Defense Analyses

Advancements in Characterizing Warhead Fragmentation Events
John Haman, Institute for Defense Analyses

Building end-to-end sustainment models for weapon systems
Han Yi, Institute for Defense Analyses

Session 7 B
Thursday, July 29
10:00 a.m. – 11:30 a.m.
Room 208

New Approaches in Design-Based Functional Analysis
Organizer and Chair: Jonathan Bradley

Simultaneous Tolerance Regions for Response Surface Designs
Aisaku Nakamura, College of Public Health, University of Kentucky

A Bayesian Functional Data Model for Surveys Collected under Informative Sampling with Application to Mortality Estimation using NHANES
Paul Parker, Department of Statistics, University of Missouri

Approximate Tolerance Regions for Simultaneous Autoregressive Models
Derek Young, Department of Statistics, University of Kentucky

Response surface models: To reduce or not to reduce?
Bryan Smucker, Department of Statistics, Miami University

Session 7 C
Thursday, July 29
10:00 a.m. – 11:30 a.m.
Room 214

Optimal Designs for Model Selection Under Penalized Estimation
Jon Stallrich, Department of Statistics, North Carolina State University

Estimating pure-error from near replicates in design of experiments
Caleb King, SAS Institute, Inc.

Applying DMAIC Principles and Data Science to Environmental Restoration
Tim Robinson, Department of Mathematics and Statistics, University of Wyoming

Session 8 A
Thursday, July 29
12:00 p.m. – 1:30 p.m.
Room 103

Applications in the DoD
Organizer and Chair: Sarah Burke

Design of Experiments for a System of Systems
Jim Simpson, JK Analytics

Developing Test Designs with Constrained Factors for Military Fast Jet Life Support Systems - A Case Study
Sarah Burke, Scientific Test and Analysis Techniques Center of Excellence (STAT COE)

Design of Experiments in Characterizing Hypersonic Flow on a Wind Tunnel Model - A Case Study
Gina Sigler, Scientific Test and Analysis Techniques Center of Excellence (STAT COE)

Session 8 B
Thursday, July 29
12:00 p.m. – 1:30 p.m.
Room 208

Bayesian Hierarchical Models for Spatially Varying Processes
Organizer and Chair: Jonathan Bradley

A Multivariate Spatial Mixture Mixed Effects Model with Application to the American Community Survey
Prof. Scott Holan, Department of Statistics, University of Missouri

Nonstationary Spatial Process Modeling using Dimension Expansion
Hou-Cheng Yang, Ph.D., U.S. Food and Drug Administration

Boundary detection for spatial processes using criteria for aggregation error
Prof. Jonathan Bradley, Department of Statistics, Florida State University

Session 8 C
Thursday, July 29
12:00 p.m. – 1:30 p.m.
Room 214

Reliability and Maintainability
Organizer and Chair: Arda Vanli, FAMU-FSU College of Engineering

Weighted multi-attribute acceptance sampling plans
Congshan Wu, Arizona State University, School of Computing, Informatics, and Decision Systems Engineering

Optimal Sampling Plan for a Multistage Production System Subject to Competing and Propagating Random Shifts
Haitao Liao, Department of Industrial Engineering, University of Arkansas

Preventive Maintenance Methods to Improve Hurricane Resilience
Arda Vanli, Department of Industrial and Manufacturing Engineering, Florida State University

An Augmented Regression Model for Tensors with Missing Values
Mostafa Reisi Gahrooei, Department of Industrial and Systems Engineering, University of Florida

Poster Session

The poster session is Tuesday, July 27, 4:30 pm to 6:00 pm. It will be a virtual session.

A Multivariate Quality Assurance Approach For Credit Card Customers and Some Features
Adnan, Mian Arif Shams
Department of Mathematics and Statistics
Bowling Green State University

Policies Review of Data Science in Industry and Government: Case of SAARC Countries
Bhandari, Dila
Nepal Commerce Campus
Tribhuvan University

SAGE: Stealthy Attack GEneration for Cyber-Physical Systems
Biehler, Michael
H. Milton School of Industrial and Systems Engineering
Georgia Institute of Technology

General Penalized Logistic Regression for Gene Selection in High-Dimensional Microarray Data Classification
Bonney, Derrick
Department of Mathematics and Statistics
Washington State University

Understanding Power Grid Network Vulnerability through the Stochastic Lens of Network Motif Evolution
Chen, Yuzhou
Department of Statistical Science
Southern Methodist University

Association-based Optimal Subpopulation Selection for Multivariate Data
Guo, Qing
Department of statistics
Virginia Polytechnic Institute and State University

Constrained Minimum Energy Designs
Huang, Chaofan
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology.

Framework for a Bayesian EWMA Control Chart using Loss Functions
Jones, Chelsea
Department of Statistical Sciences and Operations Research
Virginia Commonwealth University

A Sequence Graph Transform based Method for Monitoring Discrete Sequence Processes
Karaca, Meserret
Department of Industrial and Systems Engineering
University of Florida

HODOR: Hold-out Design for Online Randomized Experiments
Larsen, Nicholas
Department of Statistics
North Carolina State University

Neural Network Gaussian Process considering Input Uncertainty and Application to Composite Structures Assembly
Lee, Cheolhei
Grado Department of Industrial and Systems Engineering
Virginia Polytechnic Institute and State University

Robustness with Respect to Class Imbalance in Artificial Intelligence Classification Algorithms
Lian, Jiayi
Department of Statistics
Virginia Polytechnic Institute and State University

Bayesian Variable Selection Through Wavelet Neural Network For Spatio-Temporal Data
Lim, Jaehui
Department of Statistics
Florida State University

Digital twin development and maximum stress prediction for composite fuselage assembly
Lutz, Tim
Grado Department of Industrial and Systems Engineering
Virginia Polytechnic Institute and State University

Tornado Property Loss Scale: Up to $8 Billion by 2025
Mahanama, Thilini
Texas Tech University

A Hybrid Regression-Ranking Model with Application in Personalized Radiomics
Mao, Lingchao
Georgia Institute of Technology

Structural Tensor-on-Tensor Regression with Interaction Effects and its Application to a Hot Rolling Process
Miao, Huihui
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Additive Tensor Decomposition Considering Structural Data Information
Mou, Shancong
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Building a Risk Adjusted Model for Monitoring Psychopharmaceutical Prescription for Children in the Child Welfare System
Thome, Julia
Department of Biostatistics
Vanderbilt University

A Novel Statistical Data Fusion Framework for Spatial Predictive Modeling with Application in Precision Medicine.
Wang, Lujia
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

NP-ODE: Neural Process Aided Ordinary Differential Equations for Uncertainty Quantification of Finite Element Analysis
Wang, Yinan
Grado Department of Industrial and Systems Engineering
Virginia Polytechnic Institute and State University

Transparent Sequential Learning for Statistical Process Control
Xie, Xiulin
Department of Biostatistics
University of Florida

Online Sequential Monitoring of Spatio-Temporal Disease Incidence Rates
Yang, Kai
Department of Biostatistics
University of Florida

A generic online nonparametric monitoring and sampling strategy for high-dimensional heterogeneous processes
Ye, Honghan
University of Wisconsin, Madison

Nonparametric Denoising of Image Sequences
Yi, Fan
Department of Biostatistics
University of Florida

A Novel Robust Coupled CP Decomposition Method Based on Alternating Direction Method of Multipliers (ADMM)
Zhao, Meng
University of Florida

Software Demonstrations

JMP Software: New Features in v16

Wednesday, July 28, 3:00 p.m. – 3:30 p.m.

Launched in March, JMP v16 adds to the many already existing tools JMP has developed aimed at helping quality practitioners monitor, improve, and optimize processes. In this session, new platforms and tools will be shown including: Model Screening – a platform that automates the process of fitting many different statistical models comparing their performance; Sentiment Analysis – a method to mine text data and create scores quantifying the positive/negative sentiment expressed, Sample Size Explorer – a new interactive tool to determine sample size for testing and estimation, and Time Series Forecasting to simultaneously fit and compare the forecast performance of many different models.

Minitab: Statistical Tools Everyone Can Use

Wednesday, July 28, 11:30 a.m. – 12:00 p.m.

Scott Kowalski and Jennifer Atlas

Minitab empowers all parts of an organization regardless of analytical background through our comprehensive set of statistical tools for exploring data, illustrating insights, and predicting future trends. Minitab can visualize, analyze and harness the power of your data to solve your toughest challenges. Enjoy new levels of productivity and collaboration with Minitab accessible through both your desktop and on the cloud with our web app. Uncover hidden relationships between variables and visualize data interactions using the interactive graph builder. Predictive Analytics tools help to get the most out of your data: discover trends, predict patterns, and identify important factors to answer even the most challenging of questions. In this tutorial, some of the new features of Minitab will be presented.

Abstracts

Banquet Speaker

Dr. Geoff Vining
Virginia Tech

What a Long Strange Trip It's Been

This talk is a reflection on my career from making pencil lead to university professor. I would love to entitle it "An Accidental Statistician," but George Box beat me to it.

I never really planned my career; yet, in retrospect there really was a coherent direction to it, with each phase building off the prior ones. The talk's goals are to honor the people who helped me on this journey and to provide guidance to those who are still creating their own careers.

Plenary Speakers

Is Statistical Process Monitoring Out of Control?

Dr. Geoff Vining
Virginia Tech

David Banks created a great deal of controversy with his 1993 paper, "Is Industrial Statistics Out of Control?" that appeared in *Statistical Science* with discussion (pages 355-409). He was particularly critical of the state of the statistical process control community at that time.

The intent of this talk is not to create controversy but to begin a dialogue with the statistical process monitoring community. The intent is to refocus the research in this area in order for it to make significant contributions to Industry 4.0.

This talk begins with a review of Shewhart's classic 1931 book *Economic Control of Quality of Manufactured Product.* It is easy to dismiss Shewhart's work as simply a historic relic. In actuality, Shewhart's perspectives on the basic questions on the active control of any process are extremely relevant for Industry 4.0. The thought underlying Shewhart's basic procedures is quite profound, especially how he viewed the *system* of chance causes. Paramount to Shewhart was the economic impact of his methodology.

The talk then builds upon Shewhart's basic ideas to develop a new general framework for statistical process control/monitoring from a non-linear modeling perspective. This framework includes as special cases standard linear models approaches as well as univariate, multivariate, and profile monitoring schemes, since they all can be expressed in terms of standard linear models.

The proposed framework represents a very different perspective to approach the basic control/monitoring field. Ideally, such a different perspective should generate an interesting conversation with the community as it begins to adapt itself to the realities posed by Industry 4.0.

Big Data? Statistical Process Control Methods Can Help!

Peihua Qiu
University of Florida

"Big data" is a buzzword these days due to an enormous amount of data-rich applications in different industries and research projects. In practice, big data often take the form of data streams in the sense that new batches of data keep being collected over time. One fundamental research problem when analyzing big data in a given application is to monitor the underlying sequential process of the observed data to see whether it is longitudinally stable, or how its distribution changes over time. To monitor a sequential process, one major statistical tool is the statistical process control (SPC) methods, which have been developed and used mainly for monitoring production lines in the manufacturing industry during the past several decades. With many new and versatile SPC methods developed in the recent research, it is our belief that SPC can become a powerful tool for handling many big data applications that are beyond the production line monitoring. In this talk, I will introduce some recent SPC methods, and discuss their potential to solve some big data problems. Certain challenges in the interface between the current SPC research and some big data applications will also be discussed.

Statistical Science: Statistics *and* Science

Will Welch
University of British Columbia

The broad theme of the presentation will be the interplay between dimensional analysis (DA) in science and statistical modelling. While the statistical focus will be on Gaussian process (GP)99999 models for computer experiments, DA has implications for statistical models in general. DA pays attention to fundamental physical dimensions when modelling scientific and engineering systems. It goes back at least a century to Buckingham (1914), but the methodology has only recently been exploited by statisticians for design and analysis of experiments (Albrecht et al., 2013; Shen et al., 2014) and computer experiments in particular (Shen and Lin, 2018; Shen et al., 2018). The basic idea is to build models in terms of new dimensionless quantities derived from the original variables. Note that “dimension” here means a unit of measurement such as length (metres) or mass (kilograms), not the number of inputs to a system. According to the DA theory, the number of new quantities required in a model is less than the number of original inputs, hence there is dimension reduction in the statistical sense too. Specifically, Buckingham's Π theorem shows that analysis of physical units of measurement allows a system to be expressed in terms of dimensionless quantities for the output and inputs. Meinsma (2019) is an

approachable review of the key ideas. By using scientifically valid input and output variables, prediction
accuracy should improve. While the goals of DA are laudable for scientific applications in general and statistical models in particular, implementation of DA is far from straightforward. Buckingham's Π theorem does not say exactly what the derived dimensionless variables should be. Indeed, there is a combinatorially large number of possible models satisfying the conditions of the theorem. Empirical approaches for finding good derived variables will be described, and the improvements in prediction accuracy will be demonstrated. The presentation is based on the Ph.D. thesis of G. Alexi Rodriguez-Arelis.

Albrecht, M. C., Nachtsheim, C. J., Albrecht, T. A., and Cook, R. D. (2013). *Experimental Design for Engineering Dimensional Analysis*, Technometrics **55** 257 – 270.

Buckingham, E. (1914). *On Physically Similar Systems; Illustrations of the Use of Dimensional Equations*.

Meinsma, G. (2019). *Dimensional and Scaling Analysis*, SIAM Review **61** 159 – 184.

Shen, W., Davis, T., Lin, D. K. J., and Nachtsheim, C. J. (2014). *Dimensional Analysis and Its Application in Statistics*, Journal of Quality Technology **46** 185 – 198.

Presentations

Abstracts are ordered alphabetically by presenting author. Titles in ***Italic Bold*** are invited talks. Titles in **Bold** are contributed talks. Titles in *Italics* are posters.

A Multivariate Quality Assurance Approach For Credit Card Customers and Some Features

Adnan, Mian Arif Shams
Department of Mathematics and Statistics
Bowling Green State University

Credit card is an important tool for the transaction purposes (eg. buying goods online, shopping in Walmart, paying bill in restaurants) to every individual (person or institution). Customer's age, gender, identity, income category, education level, months inactive, credit limit, total revolving balance, average opening balance to buy, total transaction amount, average utilization ratio, etc are the important features of a credit card company. Attempts have made here to demonstrate several multivariate features of credit card customers and their transactions. Some multivariate quality assurance/maintenance approaches have also been developed. These methodologies can also be applied to propose the proper derivatives in share market's behaviors.

A Predictive Bayesian Approach to Sequential Time-Between-Events Monitoring

Ali, Sajid
Department of Statistics
Quaid-i-Azam University

A fundamental problem with all process monitoring techniques is the requirement of a large Phase-I data set to establish control limits and overcome estimation error. This assumption of having a large Phase-I data set is very restrictive and often problematic, especially when the sampling is expensive or not available, e.g., time-between-events (TBE) settings. Moreover, with the advancement in technology, quality practitioners are now more interested in online process monitoring. Therefore, the Bayesian methodology not only provides a natural solution for sequential and adaptive learning but also addresses the problem of a large Phase-I data set for setting up a monitoring structure. In this study, we propose Bayesian control charts for TBE assuming homogenous Poisson process. In particular, a predictive approach is adopted to introduce predictive limit control charts. Beside the Bayesian predictive Shewhart charts with dynamic control limits, a comparison of the frequentist sequential charts, designed by using the unbiased and biased estimator of the process parameter, is also a part of the present study. To assess the predictive TBE chart performance in the presence of the practitioner-to-practitioner variability, we use the average of the in-control average run length (AARL) and the standard deviation of the in-control run length (SDARL).

Concurrent Physics-Data-Driven Learning Methods for Resilient Cyber-Physical Systems

Anubi, Olugbenga Moses
Department of Electrical and Computer Engineering
Florida A&M University

The term cyber-physical system (CPS) has been coined to refer to systems with tightly coupled and integrated communication, computation and physical components. Examples include the smart grid, Intelligent traffic system, the smart water/sewer system, smart power generation, autonomous vehicles, etc. With the recent growth in the development of Internet-of-things (IoT) technology, and their application for improving the efficiency and performance of systems via added data-driven intelligence, today's critical infrastructures are essentially CPSs.

However, the cyber-threat landscape on these critical infrastructures seems to widen with increasing sophistication associated with the use of IoT despite significant benefits in terms of reliability and efficiency. The reason for this can be attributed to the advancements in information and communication technologies on the one hand and the inadequate level of security measures on legacy systems on the other hand. As a result, the chance of successful cyber-attacks has increased dramatically in the past few years. The goal of the attacker, in most cases, is to force the system to be in the wrong operational state by strategically corrupting selected process measurements. This is made possible because the physics of the system

operation is often a public knowledge and any knowledgeable attacker can figure out a strategy for corrupting the systems measurements to achieve their goal.
To address the issues, existing efforts have used one of two strategies; (a) protect a set of measurements, and (b) verify each decision variables independently. However, the high computational and deployment cost, as well as significant risk involved with these approaches, have hampered their feasibility for use in practical real-time systems. Thus, more computationally feasible, adaptive and real-time implementable resiliency strategies are needed.

In this talk, I will present some of the recent results from leveraging Machine Learning to develop algorithms and methods for the resilient operation of CPSs under adversarial attack. In order to arrive at feasible algorithms for resilient operations, there are three major issues that must be addressed. First, figure out an alternative source of information to reveal the truth of the system states even if measurements are corrupt. Second, the alternative model must be merged seamlessly with the physics-based model, and lastly, the resulting algorithm must be fast enough for real-time implementation. I will also talk about a specific examples, including how correlating readily available electricity locational marginal prices with power measurements have led to over 50% more secure algorithms over state-of-the-art methods.

Test Design Challenges in Defense Testing

Avery, Kelly
Institute for Defense Analyses

Before the DoD acquires any major new capability, that system must undergo realistic testing in its intended environment with military users. Evaluations of these systems increasingly rely on computer models and simulations (M&S) to supplement live testing in cases where testing against realistic threats is impossible, unsafe, or prohibitively expensive. It is thus crucial to thoroughly validate these M&S tools using rigorous data collection and analysis strategies to ensure their output adequately represents reality.

While classical experimental design techniques have been widely adopted across the defense community for planning live tests, gold standard computer experiment techniques from the academic literature (e.g. space filling designs and Gaussian process emulators) are much less commonly used for M&S studies. Defense testing poses unique demands, such as a heavy reliance on categorical factors and binary outcomes, extreme constraints in test conditions, small validation sets, and non-deterministic M&S outputs. There is currently no consensus on how to incorporate these demands into the existing academic framework for M&S. This talk will first provide an overview of operational testing and discuss example defense applications of, and key differences between, classical and space-filling designs. I will then present several challenges (and possible solutions) associated with implementing space-filling designs and associated analyses in the defense community.

Efficient change detection in point processes driven by data-dependent random intensities

Bhaduri, Moinak
Department of Mathematical Sciences
Bentley University

The last decade has witnessed a considerable surge in interest around change-finding within time series structures, especially with regards to certain specific features (like the mean or the variance). Similar questions are tricky to answer, even to formulate, within a point process framework where data-tracking evolves over some continuous space-time. This work offers ways to pinpoint structural breaks in such point process intensities that are influenced by both exogeneous and endogenous factors. Sequential testings are conducted through functions of trend-switched statistics and deviations within both immigrant and offspring kernels are found efficiently without sounding too many false alarms. Certain options that stay asymptotically consistent are deployed to measure the detection delay, quantified through variations of the Hausdorff metric. We will show how forecasting improves as well. Change point-based clustering tools will be offered. Examples will be chosen from COVID infection-tracking, economic announcements, and global terrorism data.

Policies Review of Data Science in Industry and Government: Case of SAARC Countries

Bhandari, Dila
Nepal Commerce Campus
Tribhuvan University

Data is the oil, Analytics is the Refinery, and Intelligence is the Gasoline which drives Growth of society. As SAARC's Big Data landscape is still growing and is facing infrastructural and policy level challenges, issues analysis method is perfect for analyzing the state of art. Information security in Nepal cyberspace is critically low and needs to address urgently. The information security policy of Nepal has serious limitations and it needs to be improved. Recently, USAID is helping ICT sector in Nepal. Users and usage aspect of big data and IOT is also not clear at policy level. Data science is a term for huge data sets having large, varied and complex structure with challenges, and difficulties in data capture, storage, analyze and visualizing for further processing for industry and government applications. Global technological advancements in the sector of information, communication and technology (ICT) have not isolated Nepalese people in adopting and moving towards digital era. The number of internet users has increased rapidly in the previous years in Nepal and South Asia. According to Nepal Telecommunication Authority (NTA) (January 15, 2019) report, total number of voice service users are 39,979,561 which is 135 % of total population (29,514,745) and total number of internet subscribers (broadband service) 17,215,980 which is about 58 % of total population. In the present globalized smart ecosystem, various suggestions of using data as a new tool for the development of the economy are still going on to be presented. As such, big data initiative is expected to stand as additional

means of data supply regardless of its complexity and statistical aspects. Reasons for growing focus on data science: Government stats agencies and barriers to private use of government data. The government of Nepal has supported liberal policies to activate the IT its applications such as big data, Internet of Things (IOT) and e-government. The goal of this paper is to describe, review and reflect on big data share the data analytics opinions and perspectives of the authors relating to the new opportunities and challenges brought forth by the big data movement and will convey an important lesson for developing countries and particularly for South Asian countries to establish policies for developing big data as a new tool for economic growth in the context of smart ecosystem environment.

SAGE: Stealthy Attack GEneration for Cyber-Physical Systems

Biehler, Michael
H. Milton School of Industrial and Systems Engineering
Georgia Institute of Technology

Cyber-physical systems (CPS) have been increasingly attacked by hackers. Recent studies have shown that CPS are especially vulnerable to insider attacks, in which case the attacker has full knowledge of the systems configuration. To better prevent such types of attacks, we need to understand how insider attacks are generated. Typically, there are three critical aspects for a successful insider attack: (i) Maximize damage, (ii) Avoid detection and (iii) Minimize the attack cost. In this paper we propose a "Stealthy Attack GEneration" (SAGE) framework by formulizing a novel optimization problem considering these three objectives and the physical constraints of the CPS. By adding small worst-case perturbations to the system, the SAGE attack can generate significant damage, while remaining undetected by the systems monitoring algorithms. The proposed methodology is evaluated on several anomaly detection algorithms. The results show that SAGE attacks can cause severe damage while staying undetected and keeping the cost of an attack low. Our method can be accessed in the supplementary material of this paper to aid researcher and practitioners in the design and development of resilient CPS and detection algorithms.

General Penalized Logistic Regression for Gene Selection in High-Dimensional Microarray Data Classification

Bonney, Derrick
Department of Mathematics and Statistics
Washington State University

High-dimensional data has become a major research area in the field of genetics, bio-informatics, and bio-statistics due to advancement of technologies. Some common issues of modeling high-dimensional gene expression data are that many of the genes may not be relevant. Also, reducing

the dimensions of the data using penalized logistic regression is one of the major challenges when there exists a high correlation among genes. High-dimensional data correspond to the situation where the number of variables is greater or larger than the number of observations. Gene selection proved to be an effective way to improve the quality and results of many classification methods. Many different methods have been proposed, however, these methods face a critical challenge in practical applications where there are high correlations among genes. The penalize logistic regression using the least absolute shrinkage selection operator (Lasso) has been criticized for being biased in gene selection. Adaptive Lasso (Alasso) was proposed to overcome the selection bias by assigning a consistent weight to each gene yet faces practical problems when choosing the type of initial weight. To address this problem, penalized logistic regression is proposed with the aim of obtaining an efficient subset of genes with high classification capabilities by combining the screening approach as a filter method and adaptive lasso with a new weight. An alternative weight in adaptive penalized logistic regression is proposed to solve this problem. We worked on existing data set and we empirically verified the proposed method performed better than other existing methods. We then used Leukemia cancer and colon cancer data set to test our proposed method. The experimental results reveal the proposed method is quite efficient and feasible and hence exhibits competitive performance in both classification accuracy and gene selection.

Boundary detection for spatial processes using criteria for aggregation error

Bradley, Jonathan
Department of Statistics
Florida State University

Spatial boundary analysis is an extremely active area of research. The inferential question of interest is often to identify rapid surface change of an unobserved latent process. Curvilinear wombling and crisp/fuzzy wombling are two major approaches that have become popular in Bayesian spatial statistics literature. These methods are limited to a single spatial scale even though data with multiple spatial scales are often accessible. Thus, we propose a multiscale representation of the directional derivative Karhunen–Loéve expansion to perform directionally based boundary detection. Taking a multiscale spatial perspective allows us, for the first time, to consider the concept of curvilinear boundary fallacy (CBF) error,which is a boundary detection analog to the ecological fallacy that is often studied in spatial change of support literature. Furthermore, we propose a directionally based multiscale curvilinear boundary error criterion to quantify CBF. We refer to this metric as the criterion for boundary aggregation error (BAGE), and use it to perform boundary detection. We illustrate the use of our model through a simulated example and an analysis of Mediterranean wind measurements.

Developing Test Designs with Constrained Factors for Military Fast Jet Life Support Systems - A Case Study

Burke, Sarah
Scientific Test and Analysis Techniques Center of Excellence (STAT COE)

Recent experiences testing one of the U.S. military's fast jet life support systems (LSS) serves as a case study to create test designs involving constrained factors. The study discusses lessons learned during unmanned LSS testing, applicable to all practitioners of design of experiments. The testing required determining a test region that included factors to model human breathing in addition to other lab settings. A comparison of government and industry laboratory tests with governing documentation is made, along with a proposal for determining an appropriate test region for tests involving human breathing as a factor.

A distribution free conditional independence test with application in causal discovery

Cai, Zhanrui
Department of Statistics and Data Science
Carnegie Mellon University

This paper is concerned with test of the conditional independence. We first establish an equivalence between the conditional independence and the mutual independence. Based on the equivalence, we propose an index to measure the conditional dependence by quantifying the mutual dependence among the transformed variables. The proposed index has several appealing properties. (a) It is distribution free since the limiting null distribution of the proposed index does not depend on the population distributions of the data. Hence the critical values can be tabulated by simulations. (b) The proposed index ranges from zero to one, and equals zero if and only if the conditional independence holds. Thus, it has nontrivial power under the alternative hypothesis. (c) It is robust to outliers and heavy- tailed data since it is invariant to conditional strictly monotone transformations. (d) It has low computational cost since it incorporates a simple closed-form expression and can be implemented in quadratic time. (e) It is insensitive to tuning parameters involved in the calculation of the proposed index. (f) The new index is applicable for multivariate random vectors as well as for discrete data. All these properties enables us to use the new index as statistical inference tools for various data. The effectiveness of the method is illustrated through extensive simulations and a real application on causal discovery.

Understanding Power Grid Network Vulnerability through the Stochastic Lens of Network Motif Evolution

Chen, Yuzhou
Department of Statistical Science
Southern Methodist University

Many cyber-physical systems can be represented as complex networks. These networks such as the internet, power grids, and supply chains, are expected to exhibit high reliability levels since their failures can lead to catastrophic cascading events. As a result, enhancing our understanding of mechanisms behind functionality of such networks is the key toward ensuring security, sustainability, and resilience of most modern critical infrastructures. While there exists a broad variety of methods for assessing reliability of networks, most existing techniques utilize only a single network topological metric. In this paper we develop a new stochastic model approach based on multiple interdependent topological measures of complex networks. The key engine behind our approach is to evaluate dynamics of multiple network motifs as descriptors of the underlying network topology and its response to adverse events. Under a framework of the gamma degradation family of models, we develop a formal statistical inference for analysis of reliability and robustness levels of a single complex network as well as for assessing differences in reliability properties exhibited by two different networks. Our studies on EU and US power grid networks indicate that the new approach delivers competitive performance, while requiring substantially less information about the underlying systems.

Fitting time series models without identification: a hierarchical sparse proximal approach

Davanloo, Sam
Department of Integrated Systems Engineering
Ohio State University

Fitting autoregressive moving average (ARMA) time series models requires model identification before parameter estimation. Model identification involves determining the order for the autoregressive and moving average components which is generally performed by visual inspection of the autocorrelation function (ACF) and partial autocorrelation (PACF) plots. In many of today's big data regime applications of time series models, however, there is a need to model one or multiple streams of data on an iterative fashion. Hence, the visual inspection step for model identification is very prohibitive. In this talk, I present a nonsmooth hierarchical sparsity inducing regularizer based on two path graphs that allows incorporating the identification into the estimation step without a need for visual inspection. I will then talk about a nonsmooth proximal optimization algorithm with convergence guarantee to solve the underlying problem. The resulting model satisfies the required stationarity and invertibility conditions for ARMA models. Finally, we present some numerical results supporting the proposed method.

The road to scaling Hotelling multivariate control charts

Dion, Francois
Dion Research LLC

Having released "Hotelling" (https://github.com/dionresearch/hotelling) as an open source python package for one and two sample Hotelling T2 tests and univariate and multivariate control charts and anomaly detection, I quickly found myself in the same scenario I had faced when building "Stemgraphic" (stemgraphic.org). With small data sets, everything just worked fine. Very large data set that do not fit in memory however were a challenge, and it seemed like every user was now dealing with tens to hundreds of gigabytes. Within a month, Stemgraphic was scaling to multiple cores and to large scale clusters.

Applying the same out of core approach used for Stemgraphic, I was able to implement large scale data set support into Hotelling version 0.5 (as of end of may, was entering beta phase with some users). The road from "in core" algorithm to full "out of core" functionality is littered with obstacles. This talk will demonstrate the use of the package and discuss some of the issues faced and the approach taken to get around them, or to solve them.

Structure of Two-Level Nonregular Designs

Edwards, David
Department of Statistical Sciences and Operations Research
Virginia Commonwealth University

Two-level fractional factorial designs are often used in screening scenarios to identify active factors. This talk investigates the block diagonal structure of the information matrix of nonregular two-level designs. This structure is appealing since estimates of parameters belonging to different diagonal submatrices are uncorrelated. As such, the covariance matrix of the least squares estimates is simplified and the number of linear dependencies is reduced. We connect the block diagonal information matrix to the parallel flats design literature and gain insights into the structure of what is estimable and/or aliased using the concept of minimal dependent sets. We show how to determine the number of parallel flats for any given design, and how to construct a design with a specified number of parallel flats using a Kronecker product construction. The usefulness of our construction method is illustrated by producing designs for estimation of the two-factor interaction model with three or more parallel flats. We also provide a fuller understanding of recently proposed group orthogonal supersaturated designs. Benefits of the block diagonal structure for analysis, including bias containment, are also discussed.

Bayesian MARS for efficient and informative nonlinear regression

Francom, Devin
Statistical Sciences Group
Los Alamos National Lab

Bayesian Multivariate Adaptive Regression Spline (BMARS) models learn the relationships between model inputs and model outputs by exploring combinations of tensor product spline basis functions in an adaptive way. In this talk, we explore how BMARS works, how it is used for uncertainty quantification tasks at Los Alamos National Laboratory, and areas of open research to improve BMARS models.

An Augmented Regression Model for Tensors with Missing Values

Gahrooei, Mostafa Reisi
Department of Industrial and Systems Engineering
University of Florida

Advancements in sensing technology created an unprecedented opportunity for developing accurate statistical models of systems with heterogeneous but complimentary sources of data. A heterogeneous set of data may contain scalars, profiles, images, and point clouds that together describe the state of a process. Models developed based on these data sets can significantly benefit process modeling, prognostics, and control. Although the existing methods, including tensor-on-tensor regression (TOT) and multiple tensor-on-tensor regression (MTOT), have shown promising results in a wide range of applications, they are only applicable to situations in which data is collected over a grid and is complete. In reality, however, the available data may not be collected over a grid (e.g., point cloud data) or may contain a large number of missing values. To address this limitation, this paper introduces a general framework that integrates tensor regression with tensor completion. Specifically, a regression model between a set of input tensors and a potentially incomplete output tensor is formulated. Next, the model parameters are estimated by minimizing a novel loss function containing a least square error and a rank penalty of the output tensor. To solve the proposed minimization problem, an efficient optimization framework that alternates between two steps is proposed. In the first step, given the model parameters, the output tensor is completed through the integration of alternating direction method of multipliers (ADMM) and augmented Lagrangian methods. In the second step, given the completed output tensor, the regression model parameters are estimated by a block coordinate descent procedure combined with an alternating least square (BCD-ALS) algorithm. Through multiple simulations and a case study, we validate the performance of our proposed method in comparison with a benchmark. The results illustrate the advantage of the proposed method over the benchmark according to the mean square prediction error criterion.

Association-based Optimal Subpopulation Selection for Multivariate Data

Guo, Qing
Department of statistics
Virginia Polytechnic Institute and State University

In the analysis of multivariate data, a useful problem is to identify a subset of observations for which the variables are strongly associated. One example is in driving safety analytics, where we may wish to identify a subset of drivers with a strong association among their driving behavior characteristics. Other interesting domains include finance, healthcare, and marketing, etc. Existing approaches, such as the Top-k method or the tau-path approach primarily relate to bivariate data and/or invoke the normality assumption. Directly adapting these methods to the multivariate framework is cumbersome. In this work, we propose a semiparametric statistical approach for the optimal selection of subpopulations based on the patterns of associations in multivariate data. The proposed method leverages the concept of general correlation coefficients to enable the optimal selection of subpopulations for a variety of association patterns. We develop efficient algorithms, Genetic Algorithm, Forward Selection, and Backward Selection for the proposed method consisting of sequential inclusion of cases into the subpopulation. We illustrate the performance of the proposed method using simulated data and also apply this method to COVID-19 data from cities in Virginia. Ten highly correlated cities are selected, and we recommend governments implement similar policies to them.

Advancements in Characterizing Warhead Fragmentation Events

Haman, John
Institute for Defense Analyses

Fragmentation analysis is a critical piece of the live fire test and evaluation (LFT&E) of lethality and vulnerability aspects of warheads. But the traditional methods for data collection are expensive and laborious. New optical tracking technology is promising to increase the fidelity of fragmentation data, and decrease the time and costs associated with data collection. However, the new data will be complex, three dimensional ‘fragmentation clouds’, possibly with a time component as well. This raises questions about how testers can effectively summarize spatial data to draw conclusions for sponsors. In this briefing, we will discuss the Bayesian spatial models that are fast and effective for characterizing the patterns in fragmentation data, along with several exploratory data analysis techniques that help us make sense of the data. Our analytic goals are to

- Produce simple statistics and visuals that help the live fire analyst compare and contrast warhead fragmentations;
- Characterize important performance attributes or confirm design/spec compliance; and

- Provide data methods that ensure higher fidelity data collection translates to higher fidelity modeling and simulation down the line.

This talk is a version of the first-step feasibility study IDA is taking – hopefully much more to come as we continue to work on this important topic.

A Multivariate Spatial Mixture Mixed Effects Model with Application to the American Community Survey

Holan, Scott
Department of Statistics
University of Missouri

Leveraging multivariate spatial dependence to improve the precision of estimates using American Community Survey data and other sample survey data has been a topic of recent interest among data-users and federal statistical agencies. One strategy is to use a multivariate spatial mixed effects model with a Gaussian observation model and latent Gaussian process model. In practice, this works well for a wide range of tabulations. Nevertheless, in situations in which the data exhibit heterogeneity within or across} geographies, and/or there is sparsity in the data, the Gaussian assumptions may be problematic and lead to underperformance. To remedy these situations, we propose a multivariate hierarchical Bayesian nonparametric mixed effects spatial mixture model to increase model flexibility. The number of clusters is chosen automatically in a data-driven manner. The effectiveness of our approach is demonstrated through a simulation study and motivating application of special tabulations for American Community Survey data.

Image-on-Scalar Partial Linear Single Index Model for Handling Heterogeneity in Imaging Studies

Huang, Chao
Department of Statistics
Florida State University

In spite of its being regarded as a powerful tool in modeling the relationship between imaging responses and multiple covariates of interest, varying coefficient models for image responses have several weaknesses: (i) the classical linear model may cause substantial prediction error when there are some constrains and nonlinear relationship between imaging responses and covariates; (ii) most existing nonlinear models (e.g., single index model) for image responses treat responses from different voxels (or vertices) as independent units, ignoring the spatial smoothness and association within the image; and (iii) the images are often collected from different centers and/phases, and the integrative analysis is challenging largely due to the substantial center/phase level heterogeneity in those image acquisitions. Our novel image-on-

scalar partial linear single index model will successfully model the nonlinear relationship between imaging responses and covariates of interest while adjusting for the imaging heterogeneity among centers/phases.

Constrained Minimum Energy Designs

Huang, Chaofan
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology.

Space-filling designs are important in computer experiments, which are critical for building a cheap surrogate model that adequately approximates an expensive computer code. Many design construction techniques in the existing literature are only applicable for rectangular bounded space, but in real world applications, the input space can often be non-rectangular because of constraints on the input variables. One solution to generate designs in a constrained space is to first generate uniformly distributed samples in the feasible region, and then use them as the candidate set to construct the designs. Sequentially Constrained Monte Carlo (SCMC) is the state-of-the-art technique for candidate generation, but it still requires large number of constraint evaluations, which is problematic especially when the constraints are expensive to evaluate. Thus, to reduce constraint evaluations and improve efficiency, we propose the Constrained Minimum Energy Design (CoMinED) that utilizes recent advances in deterministic sampling methods. Extensive simulation results on 15 benchmark problems with dimensions ranging from 2 to 13 are provided for demonstrating the improved performance of CoMinED over the existing methods.

Nonlinear profile monitoring with single index models

Iguchi, Takayuki
Department of Statistics
Florida State University

We consider change-point detection and estimation in sequences of multivariate functional observations. This setting often arises when the quality of a process is characterized by such observations, called profiles, and monitoring profiles for changes in structure can be used to ensure the stability of the process over time. We propose a nonparametric approach using a single-index model (SIM), where instead of monitoring the profile itself for a change, we monitor a statistic based on the coefficients generated by the SIM. Through simulation we show our novel method outperforms existing multivariate profile monitoring methods.

Framework for a Bayesian EWMA Control Chart using Loss Functions

Jones, Chelsea
Department of Statistical Sciences and Operations Research
Virginia Commonwealth University

In profile monitoring control charts are used to visually observe process behaviors. Often there are control charts that are best suited for varying processes. A Bayesian approach to the popular exponentially weighted moving average (EWMA) control chart is proposed to perform under different distributions. Different loss functions are considered to inform the framework of these control charts for the Poisson conjugate case. The method is assessed under a simulation study where the performance measurements are the average run length (ARL), the standard deviation of the run length (SDRL), the average time to signal (ATS), and standard deviation of time to signal (SDTS). Further assessment of the general use of the method is done via a sensitivity analysis for the control chart decision parameters, out-of-control shift sizes, and distribution hyper-parameters. Once the performance and generalization of the charts are considered, we model a count series of respiratory disease-related hospitalizations for people over 65 years old in São Paulo, Brazil. The control charts are implemented with an analysis of the hospitalization data to showcase the efficacy of our method on over-dispersed count data.

A Sequence Graph Transform based Method for Monitoring Discrete Sequence Processes

Karaca, Meserret
Department of Industrial and Systems Engineering
University of Florida

A discrete sequence is an ordered series of discrete events where an event can be described as a measured data point or an entity such as a system log and protein sequences. It is becoming customary to monitor the sequence processes for anomaly detection given that the large amounts of data are collected in these domains every day. A sequence pattern is characterized by (i) occurrences of alphabets and (ii) order of the alphabets' occurrences. Any change in either of these dimensions can be due to an anomaly. In some fields, such as cybersecurity and online banking, it is crucial to prevent any loss due to downtime or fraudulent activity. Online anomaly detection enables analysts to take preventive and curative measures as soon as an anomaly is detected. Furthermore, scalability to discrete sequences with large alphabet set is also not feasible with most of the existing methods. We consider discrete sequences in two different formats: (i) a discrete data stream: a long sequence is generated in a streaming fashion, where events get added to the same sequence over time; (ii) multiple distinct sequences: several instances of sequences generated simultaneously over time. In this study, we propose a monitoring methodology that uses sequence graph transforms (SGT). SGT is a feature extraction technique for sequences. It extracts the sequence pattern characteristics by utilizing the relative positions of alphabets in it. It is shown to be a better sequence feature extraction method, that is

more efficient for a variety of sequence problems (Ranjan, 2015). Besides, SGT does not rely on any underlying parametric distribution for sequence generation, making to robust to any scenario. Additionally, it can scale the high rate and high dimensional incoming data without exhausting the computational capacity. In this study, we bring our monitoring methodology to the experimentation using simulated and real-life datasets. Finally, we show our results in terms of accuracy and the anomaly detection time compared to existing methods such as Chi-Square monitoring and hidden Markov models.

Covariate Dependent Sparse Functional Data Analysis

Kim, Minhee
Department of Industrial and Systems Engineering
University of Wisconsin-Madison

This study proposes a method to incorporate intrinsic covariate information into sparse functional data analysis. The method aims at cases where each subject has a small number of longitudinal measurements and records intrinsic covariates. Unlike extrinsic covariates such as environmental variables which may change over time, intrinsic covariates represent the basic nature of a subject and are unvarying. One representative example is void swelling, a nuclear-specific material degradation mechanism. Void swelling is affected by many intrinsic covariates including alloy composition and material structure of the component. How to effectively model the complex relations between these covariates and the swelling process is the key to mitigating the effect of swelling and ensuring safe operation. Existing methods often do not use covariate information or only model the additive effects of extrinsic covariates. Yet, the incorporation of intrinsic covariate information into functional data analysis can significantly improve modeling and prediction performance. We first use a functional PCA to capture the functional similarities. Independent Gaussian processes then model the relations between functional principal component scores and covariates. In particular, the proposed method decomposes the variation of measurements into the variation coming from intrinsic covariates and the variation left conditioned on intrinsic covariates such that the functional PCA and Gaussian process can be conducted in a unified manner. We also develop a covariate selection algorithm based on a self-tuning variant of HMC. The proposed methodology is demonstrated using numerical experiments.

Estimating pure-error from near replicates in design of experiments

King, Caleb
SAS Institute, Inc.

In design of experiments, setting exact replicates of factor settings enables estimation of pure-error; a model-independent estimate of experimental error useful in communicating inherent system noise and testing of model lack-of-fit. Often in practice, the factor levels for replicates are

precisely measured rather than precisely set, resulting in near-replicates. This can result in inflated estimates of pure-error due to uncompensated set-point variation. In this talk, we review previous strategies for estimating pure-error from near-replicates and propose a simple alternative. We derive key analytical properties and investigate them via simulation. Finally, we illustrate the new approach with an application.

Joint work with Thomas Bzik and Peter Parker.

A Bayesian Approach for Joint Estimation for Sparse Canonical Correlation and Graphical Models

Kulkarni, Siddhesh
University of Louisville

In principle, it can be challenging to integrate data measured on same individuals occurring from different experiments and model it together to get a larger understanding of the problem. Canonical Correlation Analysis (CCA) provides a useful tool for establishing relationships between such data sets. When dealing with high dimensional datasets, Structured Sparse CCA (sSCCA) is a rapidly developing methodological area which helps to extract signal from vast amount of noise present in the data considering its structure which results in sparse direction vectors which are used to calculate CCA. There is less development in Bayesian methodology in this area. In our project we use a latent variable model, whereby using horseshoe prior, we bring in sparsity at projection matrix level, as well as at the structure level by using graphical horseshoe prior on covariance matrix to model datasets on same individuals from two experiments. We compare our results with some competing methods in a series of simulation studies.

Joint work with Jeremy Gaskins and Subhadip Pal.

Multi-task Learning with High-Dimensional Noisy Images

Kundu, Suprateek
Department of Biostatistics and Bioinformatics
Emory University

Recent medical imaging studies have given rise to distinct but inter-related datasets corresponding to multiple experimental tasks or longitudinal visits. Standard scalar-on-image regression models that fit each dataset separately are not equipped to leverage information across inter-related images, and existing multi-task learning approaches are compromised by the inability to account for the noise that is often observed in images. We propose a novel joint scalar-on-image regression framework involving wavelet-based image representations with grouped penalties that are designed to pool information across inter-related images for joint learning, and which explicitly accounts for noise in high-dimensional images via a projection-

based approach. In the presence of non-convexity arising due to noisy images, we derive non-asymptotic error bounds under non-convex as well as convex grouped penalties, even when the number of voxels increases exponentially with sample size. A projected gradient descent algorithm is used for computation, which is shown to approximate the optimal solution via well-defined non-asymptotic optimization error bounds under noisy images. Extensive simulations and application to a motivating longitudinal Alzheimer's disease study illustrate significantly improved predictive ability and greater power to detect true signals, that are simply missed by existing methods without noise correction due to the attenuation to null phenomenon.

HODOR: Hold-out Design for Online Randomized Experiments

Larsen, Nicholas
Department of Statistics
North Carolina State University

A/B tests are standard tools for estimating the average treatment effect (ATE) in online controlled experiments (OCEs), and are key to how online businesses use data to improve products and services. The majority of OCE theory makes the Stable Unit Treatment Value Assumption, which presumes the response of individual users depends only on the assigned treatment, not the treatments of others. Violations of this assumption occur when users are subjected to network interference. Standard methods for estimating the ATE typically ignore this, producing heavily biased results that limit statistical analysts' ability to increase product quality. Additionally, user covariates that are not observed, but influence both user response and network structure, also bias current ATE estimators. This fact has so far been almost completely overlooked in the network A/B testing literature. In this paper, we demonstrate that the network-influential lurking variables can heavily bias popular network clustering-based methods, thereby making them unreliable. To address this problem, we propose a two-stage design and estimation technique called HODOR: Hold-Out Design for Online Randomized experiments. The proposed method not only outperforms existing techniques, it provides reliable estimation even when the underlying network is unknown or uncertain.
Joint work with Srijan Sengupta and Jonathon Stallrich.

Neural Network Gaussian Process considering Input Uncertainty and Application to Composite Structures Assembly

Lee, Cheolhei
Grado Department of Industrial and Systems Engineering
Virginia Polytechnic Institute and State University

Application of data-driven methods for improving quality and efficiency of manufacturing processes is ubiquitous in modern industry fields, and accurate prediction for output of processes

is critical to high quality of products. Gaussian Processes have been widely used for this purpose due to its tractability and reliable uncertainty quantification property, while its shallow structure may lead to insufficient capacity of models for highly nonlinear processes. Deep Neural Network induced Gaussian Processes have expanded the capacity, yet up to now related work have assumed their inputs are noise-free, which is an inappropriate assumption for real-world cases. To achieve satisfactory quality from highly complex processes with noisy input, we propose a Neural Network Gaussian Process considering input uncertainty. Deep architecture of our regression model allows us to approximate a complex process better, and consideration of input uncertainty enables robust modeling with complete incorporation of the process uncertainty. In simulation study, we show that the proposed method can realize more accurate and stable prediction performance than other benchmark methods. At the last, we apply our approach to a composite fuselage assembly problem, which involves the highly nonlinear and anisotropic properties of composite materials and inevitable uncertainty in the assembly process.

On the characteristics of the Generalized Shiryaev-Roberts procedure for the quickest change point detection in continuous time

Li, Kexuan
Worcester Polytechnic Institute

We consider the classical Pollak-minimax quickest change-point detection problem in continuous time where the aim is to control the drift of a “live”-monitored Brownian motion. We focus on the performance of the Generalized (headstarted) Shryaev-Roberts (GSR) detection procedure set up to “watchdog” the drift of the Brownian motion. We obtain analytically a closed-form formula for the GSR procedure's Pollak's average delay to detection (given that no false alarm has yet been sounded); the formula is valid for any change-point as well as any headstart. With the aid of the formula we then show numerically that the GSR procedure with a carefully designed fixed headstart is nearly minimax-optimal when the false alarm risk is vanishingly small. This is a continuous-time equivalent of exactly the same result previously obtained in the literature for the discrete-time version of the problem.

Inference in High Dimensional Linear Measurement Error Models

Li, Mengyan
Mathematical Sciences Department
Bentley University

For a high dimensional linear model with a finite number of covariates measured with errors, we study statistical inference on the parameters associated with the error-prone covariates, and propose a new corrected decorrelated score test and a corresponding score type estimator. This work was motivated by a real data example, where both low dimensional phenotypic variables and high dimensional genotypic variables, single nucleotide polymorphisms (SNPs), are

available. One of the phenotypic variables is of clinical interest but measured with error. As is standard in the literature, the high dimensional SNPs are assumed to be measured accurately.

We show that the limiting distribution of our corrected decorrelated score test statistic is standard normal under the null hypothesis and retains power under the local alternatives around zero. We also establish the asymptotic normality of the newly proposed estimator, and hence asymptotic confidence intervals can be constructed. The finite-sample performance of the proposed inference procedure is examined through simulation studies. We further illustrate the proposed procedure via an empirical analysis of the real data example mentioned above.

Linear Hypothesis Testing in Linear Models with High Dimensional Responses

Li, Runze
Department of Statistics
Pennsylvania State University

In this paper, we propose a new projection test for linear hypotheses on regression coefficient matrices in linear models with high dimensional responses. We systematically study the theoretical properties of the proposed test. We first derive the optimal projection matrix for any given projection dimension to achieve the best power and provide an upper bound for the optimal dimension of projection matrix. We further provide insights into how to construct the optimal projection matrix. One- and two-sample mean problems can be formulated as special cases of linear hypotheses studied in this paper. We both theoretically and empirically demonstrate that the proposed test can outperform the existing ones for one- and two-sample mean problems. We conduct Monte Carlo simulation to examine the finite sample performance and illustrate the proposed test by a real data example.

Linear regression and its inference on noisy network-linked data

Li, Tianxi
Department of Statistics
University of Virginia

Linear regression on a set of observations linked by a network has been an essential tool in modeling the relationship between response and covariates with additional network data. Despite its wide range of applications in many areas, such as social sciences and health-related research, the problem has not been well-studied in statistics so far. Previous methods either lack inference tools or rely on restrictive assumptions on social effects, and usually assume that networks are observed without errors, which is too good to be true in many problems. In this talk, we propose a linear regression model with nonparametric network effects. Our model does not assume that the relational data or network structure is exactly observed; thus, the method can be provably robust to a certain level of perturbation of the network structure. We establish a set of asymptotic inference results under a general requirement of the network perturbation and then study the robustness of our method in the specific setting when the perturbation comes from random network models. We discover a phase-transition phenomenon of inference validity concerning

the network density when no prior knowledge about the network model is available, while also show the significant improvement achieved by knowing the network model. A by-product of our analysis is a rate-optimal concentration bound about subspace projection that may be of independent interest. We conduct extensive simulation studies to verify our theoretical observations, and demonstrate the advantage of our method over a few benchmarks in terms of accuracy and computational efficiency under different data-generating models. The method is then applied to adolescent network data to study gender and racial difference in social activities.

Robustness with Respect to Class Imbalance in Artificial Intelligence Classification Algorithms

Lian, Jiayi
Department of Statistics
Virginia Polytechnic Institute and State University

Artificial intelligent (AI) algorithms, such as deep learning and XGboost, are used in numerous applications including computer vision, autonomous driving, and medical diagnostics. The robustness of these AI algorithms is of great interest as inaccurate prediction could result in safety concerns and limit the adoption of AI systems. We propose a framework based on design of experiments to systematically investigate the robustness of AI classification algorithms. A robust classification algorithm is expected to have high accuracy and low variability under different application scenarios. The robustness can be affected by a wide range of factors such as the imbalance of class labels in the training dataset, the chosen prediction algorithm, the chosen dataset of the application, and a change of distribution in the training and test datasets. The poster will briefly introduce all combinations of distributions for training and testing datasets. Next, I will show how we conduct a comprehensive set of mixture experiments to collect prediction performance results. Finally, according to statistical analyses we obtained, I will provide summary of our findings and corresponding discussions to help AI practitioners better understand how various factors affect the robustness of AI classification algorithms.

Joint work with Laura Freeman, Yili Hong, and Xinwei Deng.

Optimal Sampling Plan for a Multistage Production System Subject to Competing and Propagating Random Shifts

Liao, Haitao
Department of Industrial Engineering
University of Arkansas

Sampling plans play an important role in monitoring production systems and reducing the costs related to product quality and maintenance. This paper addresses the problem of finding the optimal sampling plan for a multistage production system subject to competing and propagating random quality shifts. In particular, a serial production system with two unreliable machines that produce a product at a constant production rate is studied. It is assumed that both machines are

subject to random quality shifts due to the presence of assignable causes and can suddenly fail with increasing failure rates. If not failed, each machine operates either in its in-control state or in its out-of-control (i.e., shifted) state with different nonconforming rates. A sampling plan is implemented at the end of the production line to determine whether the system has shifted or not. If a process shift is detected, a necessary maintenance action will be initiated. The optimal sample size, sampling interval, and acceptance threshold are determined by minimizing the long-run cost rate subject to the constraints on average time to signal a true alarm, effective production rate, and system availability. A numerical example is provided to illustrate the application of the proposed sampling plan, and detailed analyses on the effects of key parameters and system constraints are also conducted.

A copula-based CUSUM scheme for monitoring auto-correlated processes

Liao, Yang-Li
Department of Information Systems, Statistics and Management Science
University of Alabama

In today's modern industry, autocorrelated processes are common due to the speed at which measurements can be obtained. The general strategy for monitoring autocorrelated processes is to design a residual-based control chart after fitting a time series model. However, this type of control chart is sensitive to the specification of the time series model, resulting in an increase of the false alarm rate when the model is misspecified. Besides autocorrelation, the validity of such parametric monitoring schemes is questionable when the underlying distribution of the process is unknown or the distributional assumptions are violated. To address the issues of model misspecification and autocorrelation, we propose a phase II copula-based cumulative sum (CUSUM) control chart with the goal of detecting location shifts in continuous autocorrelated processes with or without distributional assumptions. We then compare the average run length (ARL) performance of our new monitoring strategy to a recently published distribution-free strategy. Results suggest the performance of the proposed monitoring strategy is comparable to the distribution-free strategy, but has more capability with respect to the magnitudes of autocorrelation that can be modeled.

Bayesian Variable Selection Through Wavelet Neural Network For Spatio-Temporal Data

Lim, Jaehui
Department of Statistics
Florida State University

We introduce a Bayesian wavelet neural network (BWNN) model for variable selection for sparsely observed spatio-temporal datasets. A Bayesian variable selection method, the Spike-and-slab (SS) method, is implemented to form a node in the BWNN. Having the SS method allows BWNN to more precisely select features instead of incorporating a large number of variables (i.e. features, or basis functions and covariates in this article) which requires difficult tuning. Placing a node consists of an

SME model after the variable selecting node is another motivating component of BWNN since the SS (by itself) does not use a single set of selected features for modeling.

Univariate and multivariate techniques for root-cause analysis of quality

Linder, Henry
Oden Technologies

At Oden, we design and implement tools for quality analysis, applied in the context of continuous process manufacturing. These analyses focus on (i) systematically and reproducibly identifying the inline process contributors to bad quality; (ii) presenting these results interactively for validation and investigation by process engineers; and (iii) deployment of predictive models for real-time quality monitoring and alerting on the factory floor.

Digital twin development and maximum stress prediction for composite fuselage assembly

Lutz, Tim
Grado Department of Industrial and Systems Engineering
Virginia Polytechnic Institute and State University

Carbon fiber composites have recently entered commercial aircraft manufacturing as building materials for primary structures such as fuselage and wing sections. However, the anisotropic material properties of composite materials pose a challenge in quantifying the effects of deviations from quality targets on failure criteria. Here, we develop a finite-element model of a composite fuselage assembly that calculates residual stresses in and around rivet joints due to deviations from nominal dimensions. We generate parameter sets corresponding to feasible shape deviations via a design of experiments. We then develop a data-driven surrogate model from the numerical simulation results that quickly predicts the maximum stress for a given assembly. The predictions can be used to inform corrective measures such as fabricating custom shims for the assembly.

Tornado Property Loss Scale: Up to $8 Billion by 2025

Mahanama, Thilini
Texas Tech University

The National Oceanic and Atmospheric Administration reports that the U.S. annually sustains about 1,300 tornadoes. Currently, the Enhanced Fujita scale is used to categorize the severity of a tornado event based on a wind speed estimate. We propose the new Tornado Property Loss scale (TPL-Scale) to classify tornadoes based on associated damage costs. The dependence between the tornado affected area and the associated property losses vary strongly over time and location.

The overall tornado damage costs forecasted by a trained long short-term memory network trained on historical data might reach $8 billion over the next five years although no systematic increase in the number and cost of disasters is observed over time.
Joint work with Dimitri Volchenkov.

A Hybrid Regression-Ranking Model with Application in Personalized Radiomics

Mao, Lingchao
Georgia Institute of Technology

Radiomics is an emerging research area that builds statistical machine learning models for medical images to uncover obscure disease characteristics. However, existing radiomics research pools data of different patients to build an average model. These one-fits-all models overlook the patient-specific anatomical and disease growth differences. Although models trained with patient-specific data are expected to significantly improve accuracy, the practical challenge is that samples collected from one patient are often very limited. As a middle ground, a model that can leverage all patients' data while personalized by the information of each specific patient is highly desirable. We propose a Hybrid Regression-Ranking (HRR) model to achieve this goal, within the context of using MRI to predict the spatial distribution of Tumor Cell Density (TCD) for patients with brain cancer. The practical impact of this work is to facilitate diagnosis and treatment of brain cancer with high precision by knowing not only person-to-person but also region-to-region difference of TCD in the brain. The methodological innovation of HRR is to integrate the labeled samples from all patients, in which both MRI images and TCD measurements are available, with ordered samples for each patient, in which TCD measurements are not available but biological knowledge exists to guide the severity ordering of certain regions over others. We apply the model on real data acquired from 35 patients with glioblastoma to obtain TCD prediction maps for each patient. Data consists of a total of 139 labeled biopsy samples and 12 radiomic features extracted from six MRI contrast images from each patient. We evaluate the performance based on Mean Absolute Error, correlation, and prediction map validity.

The restricted mean time in favor of treatment

Mao, Lu
Department of Biostatistics
University of Wisconsin-Madison.

The restricted mean time in favor (RMT-IF) of treatment is a nonparametric, ICH-E9(R1)-compliant effect size estimand for complex life history data. It is defined as the net average time the treated spend in a more favorable state than the untreated over a fixed time window. It generalizes the familiar restricted mean survival time (RMST) to account for (possibly recurrent) nonfatal events in disease progression. The overall estimand admits an elegant decomposition

into stage-wise effects, with the standard RMST as a component. Via a simple re-expression, the target estimands can be estimated using plug-in Kaplan--Meier estimators for a hierarchical sequence of univariate composite events. The dynamic profile of the estimated treatment effects as a function of follow-up time can be visualized using a multilayer, cone-shaped “bouquet plot". We illustrate the proposed methods on a colon cancer trial with relapse and death as outcomes and a cardiovascular trial with recurrent hospitalizations and death as outcomes. The methodology developed is implemented in the R-package rmt, publicly available from the Comprehensive R Archive Network (CRAN).

Why are Statistical Engineers needed for Test & Evaluation? Lessons learned from my time spent in the desert, the air, windowless rooms, and sitting behind a computer.

Medlin, Rebecca
Institute for Defense Analyses

The Department of Defense (DoD) develops and acquires some of the world’s most advanced and sophisticated systems. As new technologies emerge and are incorporated into systems, OSD/DOT&E faces the challenge of ensuring that these systems undergo adequate and efficient test and evaluation (T&E) prior to operational use. Statistical engineering is a collaborative, analytical approach to problem solving that integrates statistical thinking, methods, and tools with other relevant disciplines. The statistical engineering process provides better solutions to large, unstructured, real-world problems and supports rigorous decision-making. In this talk, I will provide a survey of several projects I have been involved in during my time at IDA that highlight why I think statistical engineers are necessary for successful T&E and several of the lessons I have learned as a statistician working in the field.

Structural Tensor-on-Tensor Regression with Interaction Effects and its Application to a Hot Rolling Process

Miao, Huihui
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

This paper proposes a method of Structural Tensor-On-Tensosr Regression considering the Interaction effects (STOTI). To alleviate the curse of dimensionality and resolve the computational challenge, the STOTI method describes the specific structure of the main and interaction effect tensors indicated by the prior knowledge of the data using corresponding regularization terms on their appropriate modes. We designed an ADMM consensus algorithm to estimate these coefficient tensors. Extensive simulations and a real case study of the hot rolling process verified the superiority of the proposed method in terms of estimation and prediction accuracy.

Additive Tensor Decomposition Considering Structural Data Information

Mou, Shancong
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Tensor data with rich structural information becomes increasingly important in process modeling, monitoring, and diagnosis in manufacturing, medical and other applications. Here structural information is referred to the structural information of tensor components such as sparsity, smoothness, low-rank, and piecewise constancy. To reveal useful information from tensor data, we propose to decompose the tensor into the summation of multiple components based on their different structural information. In this paper, we provide a new definition of structural information in tensor data. We then propose an additive tensor decomposition (ATD) framework to extract useful information from tensor data. This framework specifies a high dimensional optimization problem to obtain the components with distinct structural information. An alternating direction method of multipliers (ADMM) algorithm is proposed to solve it, which is highly parallelable and thus suitable for the proposed optimization problem. Two simulation examples and a real case study in medical image analysis illustrate the versatility and effectiveness of the ATD framework.

Simultaneous Tolerance Regions for Linear Models

Nakamura, Aisaku “Isaac”
College of Public Health
University of Kentucky

A tolerance interval is a type of statistical interval that captures at least a certain proportion of the sampled population (or content) for a given confidence level. Pointwise tolerance intervals for linear models are available and widely used, however, there is no gold standard to estimate simultaneous tolerance intervals for linear models. The available simultaneous methods have various practical drawbacks, such as only being applicable locally, being excessively conservative, and being challenging to compute. One method easy to implement is the product set method introduced by Limam and Thomas [JASA, 83, 801-804 (1988)], but this is an excessively conservative approach. We find that a simple adjustment on the confidence level uniformly reduces that conservatism closer to the nominal coverage level. Extensive numerical results demonstrate the efficacy of our proposed adjustment.

Joint work with Derek S. Young.

Wavelet Methods for Very-short term Forecasting of Functional Time Series

Nystrom, Jared
Department of Operations Sciences
Air Force Institute of Technology

Space launch operations at Kennedy Space Center and Cape Canaveral Space Force Station (KSC/CCSFS) are complicated by unique requirements for near-real time determination of risk from lightning. Lightning forecast weather sensor networks produce data that are noisy, high volume, and high frequency time series for which traditional forecasting methods are often ill-suited. Current approaches result in significant residual uncertainties and consequentially may result in forecasting operational policies that are excessively conservative or inefficient. This work proposes a new methodology of wavelet-enabled semiparametric modeling to develop accurate and timely forecasts robust against chaotic functional data. Wavelets methods are first used to denoise the weather data, which is then used to estimate a single-index model for forecasting. This semiparametric technique mitigates noise of the chaotic signal while avoiding any possible distributional misspecification. Results indicate a promising technique for operationally relevant lightning prediction from chaotic sensor measurements.

A Bayesian Functional Data Model for Surveys Collected under Informative Sampling with Application to Mortality Estimation using NHANES

Parker, Paul
Department of Statistics
University of Missouri

Functional data are often extremely high-dimensional and exhibit strong dependence structures but can often prove valuable for both prediction and inference. The literature on functional data analysis is well developed; however, there has been very little work involving functional data in complex survey settings. Motivated by physical activity monitor data from the National Health and Nutrition Examination Survey (NHANES), we develop a Bayesian model for functional covariates that can properly account for the survey design. Our approach is intended for non-Gaussian data and can be applied in multivariate settings. In addition, we make use of a variety of Bayesian modeling techniques to ensure that the model is fit in a computationally efficient manner. We illustrate the value of our approach through an empirical simulation study as well as an example of mortality estimation using NHANES data.

NASA's Statistical Engineering Journey

Parker, Peter
National Aeronautics and Space Administration

The practice of statistical engineering within NASA has gained recognition and acceptance by demonstrating significant impact. The discipline's journey has gone from unknown, to rejected, to accepted, to sought after, and it has included a mixture of leaps ahead and setbacks along the

way. Currently, the state of the discipline within NASA is stable and growing, a stage worthy of accolade. However, the final destination state of fully embracing the discipline and infusing it as the way NASA routinely conducts aerospace research and develop in the execution of its missions still lies ahead. The vision is to go from a few highly-skilled practitioners, to broader workforce awareness and practice, and to go from limited applications to ubiquitous implementation. While these are lofty goals, they are worth pursuing to accelerate learning, ensure strategic resource investment, and inform rigorous decision making in NASA's missions. This presentation starts with a retrospective view of NASA's journey of introducing and implementing statistical engineering in aeronautics, space exploration, and atmospheric science that highlights milestones in technical and organizational impact, targeted educational resources, and methodological extensions that has grown its acceptance by the aerospace community. Then, we offer an assessment of the current state of practice that celebrates strengths and identifies weaknesses. In the final portion of the presentation, we propose vital elements to reach our aspired destination of greater awareness, routine application, and broadened technical and organizational impact of statistical engineering within NASA.

Co-authors: Sara Wilson and Ray Rhew

Change Point Detection Methods for Renewable Energy Integration

Pena, Malaquias
Department of Civil & Environmental Engineering
University of Connecticut

The use of change point methods to detect artificial or natural discontinuities and regime shifts in evolving weather variables have the potential to improve renewable energy management. Electric grid operators are usually concerned with abrupt changes in the amount of energy produced either by wind and solar resources as they must match demand and supply of energy at all times to avoid a blackout or other cascading problem. Integrating renewable energy into the grid has been possible by aggregating energy from multiple resources and by mixing them with non-renewable energy resources. However, as the grid becomes less centralized and more distributed into microgrids, and as the penetration of renewables increases, modern grids are much more dependent on weather variables, and thus exposed to sudden changes. These are called power ramp up or ramp down for sudden increase or decrease in energy production, respectively. In this study, we focus on the energy generated by an offshore wind farm in the New England region, and assess the performance of predicting ramp ups and ramp downs based on change point detection methods and compare it with the performance of an operational numerical weather prediction model. Both approaches provide complementary information but change point detection methods have advantages at the very short lead time.

Estimation of causal quantile effects with a binary instrumental variable and censored data

Peng, Limin
Department of Biostatistics
Emory University

In this work, we study a new binary IV framework with randomly censored outcomes where we propose to quantify the causal treatment effect by the concept of complier quantile causal effect (CQCE). The CQCE is identifiable under weaker conditions than the complier average causal effect when outcomes are subject to censoring, and it can provide useful insight into the dynamics of the causal treatment effect. Employing the special characteristic of the binary IV and adapting the principle of conditional score, we uncover a simple weighting scheme that can be incorporated into the standard censored quantile regression procedure to estimate CQCE. We develop robust nonparametric estimation of the derived weights in the first stage, which permits stable implementation of the second stage estimation based on existing software. We establish rigorous asymptotic properties for the proposed estimator, and confirm its validity and satisfactory finite-sample performance via extensive simulations. The proposed method is applied to a bone marrow transplant dataset to evaluate the causal effect of rituximab in diffuse large B-cell lymphoma patients.

A distribution-free joint monitoring scheme for location and scale using individual observations

Perry, Marcus
Department of Information Systems, Statistics and Management Science
University of Alabama

Recent advances in data acquisition and storage technologies have permitted the rapid col- lection of data over time at a relatively low cost. The implication of these advances to mod- ern quality engineering is that many of today's processes produce samples of individual observations that are grossly non-normal and, potentially, very highly autocorrelated. Consequently, the typical assumptions required by traditional control charting strategies for individual observations are not likely to be met by today's more modern processes. This presents a significant challenge for today's quality engineer practitioner, particularly when the false alarm rate of the monitoring strategy should be adequately controlled. In this effort, we propose a new joint monitoring scheme for location and scale using individual observations that relaxes some of the assumptions that limit the use of traditional control charts in today's practice. In addition, the proposed method is extremely practitioner- friendly and easy to implement. We compare performances of our new scheme to the commonly-used individuals and moving range control charts. Results suggest the proposed scheme provides an effective and robust means to jointly monitor the kind of processes most prevalent in today's modern industry. We demonstrate our method using open-source data available from a selective laser melting (SLM) process, where the detection of hot spots at a given location on a manufactured part was of interest.

Statistical Engineering and Wicked Problems?

Perry, Marcus
Department of Information Systems, Statistics and Management Science
University of Alabama

Statistical Engineering is a holistic approach to solving large, complex and unstructured (LCU) problems. Currently, most solutions to complex problems are often found and maintained by design strategies that are rooted in the Theory of Wicked Problems. In this talk, we review the Theory of Wicked Problems and discuss its compatibility with the Statistical Engineering discipline. Outlining the consistencies and inconsistencies between Statistical Engineering and the Theory of Wicked Problems can help to better distinguish and further define the former as a legitimate discipline.

On the performance of the generalized Shiryaev-Roberts control chart in continuous time

Polunchenko, Aleksey S.
Department of Mathematical Sciences
Binghamton University

The topic of interest is the performance of the Generalized Shiryaev-Roberts (GSR) control chart in continuous time, where the goal is to detect a possible onset of a drift in a standard Brownian motion observed live. We derive analytically and in a closed-form all of the relevant performance characteristics of the chart. By virtue of the obtained performance formulae we show numerically that the GSR chart with a carefully designed headstart is far superior to such mainstream charts as CUSUM and EWMA. More importantly, the Fast Initial Response feature exhibited by the headstarted GSR chart makes the latter not only better than the mainstream charts, but nearly the best one can do overall for a fixed in-control Average Run Length level. This is a stronger conclusion than that previously reached about CUSUM in the seminal 1982 Technometrics paper by J.M. Lucas and R.B. Crosier.

Monitoring processes with a large number of variables

Potcner, Kevin
SAS Institute Inc.

As the amount of data that can be collected on processes grows, understanding and improving these processes has moved beyond single variable control charts and capability analyses. Today's quality practitioners need to add analysis tools to their toolkit that are designed to understand processes and systems when a large number of variables are required. The presenter will illustrate a variety of these tools including Process Screening - a tool that can be used to monitor the performance of a process containing many different process variables, Multivariable control charts and capability analyses, and Response Screening - a technique designed to determine statistical significance across a large of number of comparisons.

Using BART for Multiobjective Optimization of Noisy Multiple Objectives

Pratola, Matthew
Department of Statistics
Ohio State University

Techniques to reduce the energy burden of an Industry 4.0 ecosystem often require solving a multiobjective optimization problem. However, collecting experimental data can often be either expensive or time-consuming. In such cases, statistical methods can be helpful. This article proposes Pareto Front (PF) and Pareto Set (PS) estimation methods using Bayesian Additive Regression Trees (BART), which is a non-parametric model whose assumptions are typically less restrictive than popular alternatives, such as Gaussian Processes. The performance of our BART-based method is compared to a GP-based method using analytic test functions, demonstrating convincing advantages. Finally, our BART-based methodology is applied to a motivating Industry 4.0 engineering problem.

Co-authors: Akira Horiguchi, Thomas J. Santner, and Ying Sun

Applying DMAIC Principles and Data Science to Environmental Restoration

Robinson, Tim
Department of Mathematics and Statistics
University of Wyoming

The decade 2021-2030 has been declared the decade of ecosystem restoration by the United Nations. It is estimated that global restoration of degraded land due to industrial and agricultural impacts could cost in excess of $12 trillion U.S. - roughly 2/3 of the U.S. GDP. Ecosystem and landscape restoration projects will under-deliver without strategic management approaches which link ecological science, systems modeling, survey design and data science. Many of the principles from the DMAIC quality improvement process can be utilized for the efficient design of restoration monitoring programs. The Prairie Pothole Region (PPR) of North American encompasses 715,000 km2 of critical habitat for waterfowl, migratory songbirds and pollinators. In many areas of the PPR, nearly 90% of the habitat has been degraded due to industry and agriculture. This session will describe the application of DMAIC principles to a habitat restoration project within the PPR.

Distortion Modeling and Compensation Across Materials and Processes in Laser-Based Additive Manufacturing Systems via Bayesian Neural Networks

Sabbaghi, Arman
Department of Statistics
Purdue University

Abstract: Distortion in laser-based additive manufacturing (LBAM) systems is a critical quality control issue that is known to be material- and process-dependent. One key challenge in LBAM systems is learning distortion models for new materials and processes given past experiments on distinct materials and processes. We present a Bayesian neural network methodology to transfer distortion models across different materials and processes in an LBAM system, and create compensation plans to reduce distortions in future additively manufactured products, thereby leveraging past experiments to learn distortion models for new materials and processes. We demonstrate the utility of this method with case studies on disks that are additively manufactured using Ti-6Al-4V and 316L stainless steel.

Joint work with Jack Francis, M. Ravi Shankar, Morteza Ghasri-Khouzani, and Linkan Bian.

Statistical methods applied to monitoring growth in shrimp farms

Sánchez, Ismael
Department of Industrial and Systems Engineering
Universidad de Piura

Shrimp is one of the most valuable traded marine products in the world. This work presents a methodology for the statistical monitoring of shrimp production in an aquaculture system. It will allow a faster and better inspection of the process, allowing you to take better decisions.

The design of control charts for shrimp growth presents several challenges, the main ones being the presence of trend and autocorrelation. Several control charts are proposed, the most efficient being a chart based on the residuals of a second order autoregression whose parameters are time varying. A control chart with less computational complexity is also provided. It is simply based on growth rates and works well at detecting negative weight changes. Shrimp farms usually have hundreds of ponds. Since the proposed control charts allow for automated monitoring, production technicians would only have to worry about those ponds that have been signaled abnormal by the control charts. This frees up time for these specialists, which can be used to focus on more relevant issues.

We apply the methodology to a large-scale fish farm in northern Peru.

Joint work with Isabel González.

Bayesian Sequential Monitoring of Density Estimates

Shamp, Wright
Johnson and Johnson

In this paper we consider sequentially estimating the density of univariate data. We utilize Polya trees to develop a statistical process control (SPC) methodology. Our proposed methodology monitors the distribution of observed data and detect when the generating density differs from the in-control standard. We also propose an approximation that merges the probability mass of multiple data points being a change point in the proposed methodology to curb computational complexity while maintaining accuracy in detecting a change point.

Experimental Design with a Functional Response

Scherer, Matthew
Department of Operational Sciences
Air Force Institute of Technology

The growth of sensor streamed data in recent years increases the demand for an analytical technique to properly address data measured continuously. The design and analysis of experiments (DOE) of U.S. Air Force assets are based off of sensor streamed data. Functional data analysis (FDA) is an approach of analyzing data existing over a continuum. This research aids in filling the intersection of FDA and DOE by examining a case study of an experimental design with a functional response in addition to insight on software capabilities in FDA. The case study considers a functional linear model of a whole-plot from a split-plot experimental design compared to multivariate methods and an approximated functional linear model. Initial results indicate no significant main effects were detected in the case study using FDA. However, a comparison between the different methodologies indicate similar behaviors for main effect estimates. An examination of software packages reveals the R software as most compatible with FDA methodology. Recommendations include another case study evaluation of FDA and future work in alignment of response curves.

Semi-Supervised Clustered Multi-Task Learning (SSC-MTL) for modeling connected systems

Seyedsalehi, Sajjad
Department of Industrial & Operations Engineering
University of Michigan

In multitask learning (MTL), multiple learning problems(e.g. regressions)will be solved together to improve performance. Away to enforce MTL is by imposing a cluster structure among tasks. Here we are proposing a new approach named Semi-Supervised Clustered Multi-Task Learning(SSC-MTL) that benefits the tasks with known cluster assignment. One application

is for modeling connected systems (with unknown cluster assignment) in the presence of experimental systems(with known cluster assignment). Our proposed approach uses the k-means algorithm concept and a specific regularization to share information between tasks with known cluster assignment and those with unknown cluster assignment.

Inverse Problems in Materials Characterization

Shanbhag, Sachin
Department of Scientific Computing
Florida State University

Rheology is extraordinarily sensitive to polymer structure. Due to its sensitivity and ease of measurement it is a promising tool for characterization. This involves inverting computational models that take the structure and composition of a polymer mixture as input, and predict the rheology as output. Using a Bayesian framework, we pose the inverse problem of material characterization as a sampling problem that can be explored using Markov chain Monte Carlo. However, evaluation of computational model is expensive, and presents a key bottleneck. We describe recent developments in using surrogate models based on Gaussian processes that make the inverse problem more amenable.

Co-authors: Pankaj Chouhan, Arsia Takeh

Design of Experiments in Characterizing Hypersonic Flow on a Wind Tunnel Model - A Case Study

Sigler, Gina
Scientific Test and Analysis Techniques Center of Excellence (STAT COE)

While wind tunnel testing is very useful for collecting data at representative flight conditions, it gives rise to several unique statistical difficulties. The non-continuous blowdown hypersonic wind tunnel used in this study remains fixed on certain factors during the short run time, requiring specific analytical techniques to account for the lack of randomization. The short duration run times coupled with relatively lengthy between-run preparation time means that tunnel time is at a premium and test designs must be efficient. This case study shows that a design of experiments methodology can efficiently and effectively account for these constraints while exploring the relationship between factors and responses of interest in the system. The goal of this wind tunnel test was to gather data to validate high-fidelity computational models for an inward turning inlet and gain a better understanding of the fundamental flow features that govern inlet operability. Advancements in high-fidelity computational models allow the operability and performance of an inward turning hypersonic inlet to be fully modeled, enabling increased vehicle capability. This work will be critical to developing shorter, lighter, and more maneuverable hypersonic systems that will be able to fly at increased ranges.

Design of Experiments for a System of Systems

Simpson, Jim
JK Analytics

What happens when a test team is confronted with the challenge of planning a test involving several systems that must be used operationally as integrated systems of systems (SoS)? In this discussion, we use a case study approach to offer the engineer/analyst a practical approach to testing a system of systems. The test process emphasis on the design and analysis phases primarily but we also address the planning and execution phases. A notional Navy information collection and fusion system serves as the example. The procedure uses a statistical engineering approach to confront various decisional aspects of the test design, data collection and data analysis. We consider two scenarios: 1) all factors are easy to change, and 2) some factors are hard to change. Competing designs are developed for each scenario and compared, leading to options for the user. Analysis is demonstrated using simulated data and the analysis methods are described along with recommendations. The outcome is a proposed strategy for design and analysis while encouraging follow-on research.

This is joint work with Jim Wisnowski.

Response surface models: To reduce or not to reduce?

Smucker, Bryan
Department of Statistics
Miami University

In classical response surface methodology, the optimization step uses a small number of important factors. However, in practice, experimenters sometimes fit a second-order model without previous experimentation. In this case, the true model is uncertain and the full model may overfit. Here, we use an extensive simulation to evaluate several analysis strategies in terms of their optimum locating ability, and use both simulation and published experiments to evaluate their general prediction facility. We consider traditional (reducing via p-values; forward selection), regularization (LASSO; Gauss-LASSO), and Bayesian analysis methods.

Optimal Designs for Model Selection Under Penalized Estimation

Stallrich, Jon
Department of Statistics
North Carolina State University

The analysis of a screening experiment often focuses more on model selection than estimation so screening design criteria should be based on the desired model selection procedure. One such approach is via penalized estimation like the LASSO and Dantzig selector that shrink estimates

to 0. In this talk, we introduce a new design theory for the LASSO that promotes desirable model selection properties. The theory resembles local and Bayesian optimality approaches for nonlinear linear models, but often fixes only the sign of the model parameters, rather than their specific values. The design measures of interest involve probabilities of either successful model or sign recovery, the latter being more stringent. When sign effects are correctly guessed, the theory establishes the superiority of constrained, positive Var(s) optimal supersaturated designs over conventional supersaturated designs in terms of model selection. Otherwise, the theory establishes that heuristic orthogonality measures (e.g., E(s2) and UE(s2)), generate equivalently powerful designs. Finally, we discuss criteria based on the thresholded lasso that describes behavior of the LASSO solution path for a design.

Optimizing the Online User Experience: Design and Analysis of A/B Tests with Bayesian Comparative Probability Metrics

Stevens, Nathaniel T.
Department of Statistics and Actuarial Science
University of Waterloo

As a means to remain competitive, many tech companies are using online controlled experiments (A/B tests) to improve their products, services, and customer experience. Companies like Google, Netflix, Microsoft and Facebook run tens of thousands of experiments, engaging millions of users, each year. Such experiments are typically used to decide whether a particular product variant outperforms one or more alternatives. Relative to traditional DOE applications, the cost of experimental units is much lower and the ease with which data are collected is much greater. This translates into enormous sample sizes that often produce statistically significant p-values, independent of whether the difference between variants is practically meaningful. In this talk we propose the use of comparative probability metrics as an estimation-based alternative to traditional two-sample hypothesis testing. The proposed Bayesian methodology provides a flexible and intuitive means of quantifying the similarity or superiority of one variant relative to another, while accounting for the magnitude of a practically inconsequential difference. A methodology for sample size determination will also be discussed.

Joint work with Luke Hagar.

Debiasing Multidimensional Scaling

Sun, Qiang
Department of Statistics
University of Toronto

Classical multidimensional scaling is an important tool for data reduction in many applications. It takes in a distance matrix and outputs low-dimensional embedded samples such that the

pairwise distances between the original data points can be preserved, when treating them as deterministic points. However, data are often noisy in practice. In such case, the quality of embedded samples produced by classical multidimensional scaling starts to break down, when either the ambient dimensionality or the noise variance gets larger. This motivates us to debias the classical multidimensional scaling procedure by applying a nonlinear shrinkage to the sample eigenvalues. The nonlinear transformation is determined by sample size, the ambient dimensionality, and the moments of noise. As an application, we consider the problem of clustering high-dimensional noisy data. We show that the debiased multidimensional scaling followed by various clustering algorithms can achieve exact recovery, i.e., all the cluster labels can be recovered correctly with probability tending to one. Numerical studies lend strong support to our proposed methodology.

A Tree-based Federated Learning Approach for Personalized Treatment Effect Estimation from Heterogeneous Data Sources

Tang, Lu
Department of Biostatistics
University of Pittsburgh

Federated learning is an appealing framework for analyzing sensitive data from distributed health data networks. Under this framework, data partners at local sites collaboratively build an analytical model under the orchestration of a coordinating site, while keeping the data decentralized. While integrating information from multiple sources may boost statistical efficiency, existing federated learning methods mainly assume data across sites are homogeneous samples of the global population, failing to properly account for the extra variability across sites in estimation and inference. Drawing on a multi-hospital electronic health records network, we develop an efficient and interpretable tree-based ensemble of personalized treatment effect estimators to join results across hospital sites, while actively modeling for the heterogeneity in data sources through site partitioning. The efficiency of this approach is demonstrated by a study of causal effects of oxygen saturation on hospital mortality and backed up by comprehensive numerical results.

Building a Risk Adjusted Model for Monitoring Psychopharmaceutical Prescription for Children in the Child Welfare System

Thome, Julia
Department of Biostatistics
Vanderbilt University

Children in the child welfare system are more likely to receive psychotropic medication prescriptions than children in the general population. This session will detail the use of

administrative- and prescription-level child data and Administration on Children, Youth and Families guidelines to quantify variability in prescribing practices among prescribers for child welfare. These data were used to build a risk adjusted model that accounts for case complexity and estimates the adjusted probability of potentially inappropriate psychotropic prescriptions (i.e., red-flagged prescriptions). It is vital to monitor psychotropic prescriptions for children in the child welfare system. Quantifying variability in prescribing practices among prescribers for these children can be used to guide oversight and identify areas of needed support and intervention. The goal of this model is to promote safe and effective prescribing in this vulnerable population.

Nonparametric multivariate profile monitoring using regression trees

Timme, Daniel
Department of Statistics
Florida State University

Monitoring profiles for changes in the behavior can be used to validate whether the process is operating under normal conditions over time. Change-point detection and estimation in sequences of multivariate functional observations is a common method utilized in monitoring such profiles. A nonparametric method utilizing Classification and Regression Trees (CART) to build a sequence of regression trees is proposed which makes use of the Kolmogorov-Smirnov statistic to monitor profile behavior. Our novel method compares favorably to existing methods in the literature.

Additive and projection pursuit models with Gaussian process regression

Tuo, Rui
Department of Industrial and Systems Engineering
Texas A&M University

A primary goal of computer experiments is to reconstruct the function given by the computer code via scattered evaluations. Traditional isotropic Gaussian process models suffer from curse of dimensionality, when the input dimension is high. Gaussian process models with additive kernels are scalable to dimensionality, but they are very restrictive as they only work for additive functions. In this work, we consider a projection pursuit model, in which the nonparametric part is driven by an additive Gaussian process regression. The dimension of the additive function is chosen to be higher than the original input dimension. We show that this dimension expansion can help approximate more complex functions. A gradient descent algorithm is proposed to maximize the likelihood function. Simulation studies show that the proposed method outperforms the traditional Gaussian process models.

Sequentially monitoring nonlinear profiles using a Gaussian process model with heteroscedasticity

Quevedo, A. Valeria
Department of Industrial and Systems Engineering
Universidad de Piura

There is extensive research in the monitoring of a process whose characteristics are represented in functional forms such as profiles. Most of the current techniques require the entire data from each profile to determine the process state. We study the use of a Shewhart chart based on a Gaussian process regression model with heteroscedasticity to monitor profiles sequentially while they are being developed. The central line of the chart is the predictive mean, and its control limits are based on the prediction band. The advantage is that we do not have to wait until a profile ends to make process corrections. Our results from the simulation and case study indicate that our method is effective.

Joint work with Geoffrey Vining.

Preventive Maintenance Methods to Improve Hurricane Resilience

Vanli, Arda
Department of Industrial and Manufacturing Engineering
Florida A&M University – Florida State University College of Engineering

This paper presents a time-dependent reliability modeling and optimal maintenance planning approach for residential buildings subject to hurricane winds. A gamma process is employed to model the stochastic degradation of building components. The time-dependent fragility function and the failure probability of the roof system obtained from the degradation model are used to determine the retrofit timing of components that optimize an average cost per cycle time. The method is illustrated on determining the retrofit times of roof-to-wall connections based on actual component failure and hurricane wind load data. Effectiveness of the optimal retrofit age is compared to a replacement based on time-invariant failure probabilities. A simulation approach is presented to quantify the effect of wind speed uncertainty on maintenance cost.

Two-level orthogonal screening designs with 80, 96 and 112 runs: Construction and evaluation

Vazquez, Alan R.
Department of Statistics
University of California, Los Angeles

Due to recent advances in the development of laboratory equipment, large screening experiments can now be conducted to study the joint impact of up to a few dozen factors. While much is

known about orthogonal designs involving 64 and 128 runs, there is a lack of literature on screening designs with intermediate run sizes. In this presentation, we therefore construct screening designs with 80, 96 and 112 runs, which allow the main effects to be estimated independently from the two-factor interactions and limit the aliasing among the interactions. Our construction procedure concatenates two smaller equally-sized orthogonal arrays of strength three. We motivate our work using a 14-factor tuberculosis inhibition experiment and compare our new designs with alternatives from the literature using simulations.

Joint work with Prof. Peter Goos and Dr. Eric Schoen from the University of Leuven, Belgium.

A Screening Algorithm for Cross-Validated Kernel Support Vector Machines

Wang, Boxiang
Department of Statistics and Actuarial Science
University of Iowa

The kernel support vector machine (SVM) is recognized as one of the most competitive classifiers and a flagship method in data science. However, the kernel SVM has limited applications in large-scale data because of its intricate computational difficulties. In this work, we introduce a novel screening method for computing the SVM, without sacrifice of the accuracy. We strategize the screening approach to identify a subset of coefficients that are inactive in both training and leave-one-out cross-validated data, and then apply a recently proposed magic SVM algorithm to simultaneously compute the active coefficients as well as leave-one-out
cross validation error. Consequently, our procedure directly yields the tuned SVM for practical use. We implement our algorithm in a publicly available R package. With simulated and real data examples, we demonstrate that our proposal is much faster than the two state-of-the-art SVM solvers: R packages kernlab and e1071.

Multi-printer Co-learning of Kinematics-Induced Variations for Inter-connected Extrusion-based Additive Manufacturing

Wang, Hui
Department of Industrial and Manufacturing Engineering
Florida State University

For extrusion-based additive manufacturing, the inconsistent amount of deposited material along with the printing path greatly affects printed material distribution, causing infill non-uniformity and defect. The quality problems are caused by kinematics variations such as speed and acceleration. However, state-of-the-art research on quality control for such variation patterns is very limited. This talk presents a mathematical model that quantifies the printing line variations induced by extruder kinematics characteristics. The model can help predict product quality and potential printing process optimization. In addition, it also provides a way of sharing knowledge between printers to help co-learn the quality problems. We proposed an informative-priors-based

transfer-learning algorithm to improve the quality prediction model for a 3D printer with limited historical data by leveraging the data from inter-connected 3D printers that have more sample data shared over a cloud. A case study based on experiments validated the effectiveness of the proposed methods. The results can enable cloud-based additive manufacturing systems with improved reconfigurability and reduced calibration time.

A Novel Statistical Data Fusion Framework for Spatial Predictive Modeling with Application in Precision Medicine.

Wang, Lujia
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

The automated capability of generating spatial prediction for a variable of interest is desirable in various science and engineering domains. Take Precision Medicine of cancer as an example, in which the goal is to match patients with treatments based on molecular markers identified in each patient's tumor. A substantial challenge, however, is that the molecular markers can vary significantly at different spatial locations of a tumor. If this spatial distribution could be predicted, the precision of cancer treatment could be greatly improved by adapting treatment to the spatial molecular heterogeneity. This is a challenging task because no technology is available to measure the molecular markers at each spatial location within a tumor. Biopsy samples provide direct measurement, but they are scarce/local. Imaging, such as MRI, is global, but it only provides proxy/indirect measurement. Also available are mechanistic models or domain knowledge, which are often approximate or incomplete. This work proposes a novel statistical machine learning framework to fuse the three sources of data/information to generate spatial prediction, namely the knowledge-infused global-local data fusion (KGL) model. A novel mathematical formulation is proposed and solved with theoretical study. We present a real-data application of predicting the spatial distribution of Tumor Cell Density (TCD)—an important molecular marker for brain cancer. A total of 82 biopsy samples were acquired from 18 patients with glioblastoma, together with 6 MRI contrast images from each patient and biological knowledge encoded by a PDE simulator-based mechanistic model called Proliferation-Invasion (PI). KGL achieved the highest prediction accuracy and minimum prediction uncertainty compared with a variety of competing methods.

NP-ODE: Neural Process Aided Ordinary Differential Equations for Uncertainty Quantification of Finite Element Analysis

Wang, Yinan
Grado Department of Industrial and Systems Engineering
Virginia Polytechnic Institute and State University

Tribocorrosion is a material degradation process involving both mechanical wear and corrosion of the material, which jeopardizes a material's long-term sustainability and structural integrity. Tribocorrosion analysis is very important for ensuring quality and reliability of various materials. Finite Element Analysis (FEA) has been widely used to investigate the effects of a material's mechanical and electrochemical properties on their tribocorrosion behavior. Despite its strength and accuracy, FEA usually has two limitations: (i) running high-fidelity FEA often requires high computational cost and consumes a large amount of time; (ii) FEA is a deterministic method that is insufficient for uncertainty quantification when modeling complex systems with various types of uncertainties. In this work, a physics-informed data-driven stochastic surrogate model, named Neural Process Aided Ordinary Differential Equation (NP-ODE), is proposed to model the FEA simulations and capture both input and output uncertainties. To validate the advantages of the proposed NP-ODE, we conduct experiments on both the simulation data generated from a given ordinary differential equation and the data collected from a real FEA platform for tribocorrosion. The results show that the proposed NP-ODE outperforms benchmark methods. The NP-ODE method realizes the smallest predictive error as well as generating the most reasonable confidence intervals with the best coverage on testing data points.

On the Simulation of Autocorrelated Processes

Wang, Zhi
Bayer Crop Science

Simulation studies are computer experiments using nominal values to create data by pseudo-random sampling from statistical models. The process of generating data allows one to assess the performance and understand the behavior of statistical methods in relation to a known truth. Such evaluation and observation often cannot be conducted by the real data alone, and this ensures the necessity of simulation. There have been considerable efforts in recent years devoted to the development of methods that are able to reflect the complex situation encountered in practice for dealing with autocorrelated data sequences instead of assuming independence. Consequently, designing a reliable and flexible simulation procedure for the evaluation of these methods is of paramount interest. This paper proposes the generalization of Bahadur representation and employs it to simulate autocorrelated data sequences from both discrete and continuous distributions. The utilization of the proposed simulation approach permits the diversity of the process distribution and guarantees the structure and order of autocorrelation are fully customizable. Real-world application and comparison studies for selected distributions are conducted to show the superior performance and the broad applicability of the proposed method.

Bootstrapping quantiles of degradation data

Warr, Richard
Department of Statistics
Brigham Young University

Finding bootstrap quantiles of data is often slow with a significant amount of Monte Caro error. We present an improved technique for calculating bootstrap quantiles. This convolutional technique utilizes the fast Fourier transform. Our approach can provide exact estimates, or at a minimum mathematical bounds. For most problems, the computation time can be dramatically less than that of Monte Carlo resampling. We apply our method to computing quantiles of degradation data and compare the results with a few competing methods in the literature.

Multi-printer Co-learning of Kinematics-Induced Variations for Inter-connected Extrusion-based Additive Manufacturing

Wei, An-Tsun
Department of Industrial and Manufacturing Engineering
Florida State University

For extrusion-based additive manufacturing, the inconsistent amount of deposited material along with the printing path greatly affects printed material distribution, causing infill non-uniformity and defect. The quality problems are caused by kinematics variations such as speed and acceleration. However, state-of-the-art research on quality control for such variation patterns is very limited. This talk presents a mathematical model that quantifies the printing line variations induced by extruder kinematics characteristics. The model can help predict product quality and potential printing process optimization. In addition, it also provides a way of sharing knowledge between printers to help co-learn the quality problems. We proposed an informative-priors-based transfer-learning algorithm to improve the quality prediction model for a 3D printer with limited historical data by leveraging the data from inter-connected 3D printers that have more sample data shared over a cloud. A case study based on experiments validated the effectiveness of the proposed methods. The results can enable cloud-based additive manufacturing systems with improved reconfigurability and reduced calibration time.

Spectral clustering via adaptive layer aggregation for multi-layer networks

Weng, Haolei
Department of Statistics and Probability
Michigan State University

One of the fundamental problems in network analysis is detecting community structure in multi-layer networks, of which each layer represents one type of edge information among the nodes. We propose integrative spectral clustering approaches based on effective convex layer aggregations. Our aggregation methods are strongly motivated by a delicate asymptotic analysis

of the spectral embedding of weighted adjacency matrices and the downstream k-means clustering, in a challenging regime where community detection consistency is impossible. In fact, the methods are shown to estimate the optimal convex aggregation, which minimizes the mis-clustering error under some specialized multi-layer network models. Our analysis further suggests that clustering using Gaussian mixture models is generally superior to the commonly used k-means in spectral clustering. Extensive numerical studies demonstrate that our adaptive aggregation techniques, together with Gaussian mixture model clustering, make the new spectral clustering remarkably competitive compared to several popularly used methods.

Wavelets on Graphs

Williams, Jeff
Department of Operations Sciences
Air Force Institute of Technology

Graphs provide structure and context to data, both of which are relevant to the analysis and subsequent insight of information flow for social networks, transportation networks, and physical structures. Graph signal processing (GSP) is a collection of tools elevated from signal processing and used to analyze a variety of abstract and physical networks. This research elevates the idea of data neighborhoods to graphs to improve existing GSP wavelet thresholding tools for smoothing and approximating graph signals. Results indicate that data neighborhoods formed based on graph orientation, rather than spatial orientation, considerably improve the effectiveness of GSP tools, such as graph wavelet analysis, and ultimately understanding of networks and respective data flow.

Finding the source of grandma's chili: investigative text mining

Wise, Scott Lee
SAS Institute, Inc.

The power of using Text Mining is a great tool in investigating all kinds of unstructured text that commonly resides in our collected data. From notes captured on warranty issues, lab testing/experimental comments, to even looking at food recipes, this new method opens a lot of opportunity to better understand our world. In this presentation, we will show how to us the latest text analytic methods to help solve a family mystery as to the regional source of my Grandma's delicious chili recipe. Along the way we will see how to use text mining to create leading terms and phrase lists and word cloud reports. Then we will utilize the resulting document term matrix to perform topic analysis (via latent class analysis clustering) that will enable us to find a solution to our question. You will be left with an understanding of the powerful text mining approaches that you can add to your own toolbox and start solving your own text data challenges!

Weighted multi-attribute acceptance sampling plans

Wu, Congshan
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University

Acceptance sampling plan plays an important role in the field of industrial statistics and has wide applications in quality control, such as food classification management, epidemic control and so on. In this talk, products with weighted multiple attributes are considered. Based on different acceptance rules, two new acceptance sampling plans are proposed to satisfy real quality control requirements. The OC functions for new sampling plans are constructed by using finite Markov chain imbedding approach. In addition, the optimal parameters for sampling plans can be determined by nonlinear optimization models. Finally, results and discussions are given to demonstrate the effectiveness of the proposed acceptance sampling plans.

Transparent Sequential Learning for Statistical Process Control

Xie, Xiulin
Department of Biostatistics
University of Florida

Machine learning methods have been widely used in different applications, including process control and monitoring. For handling statistical process control (SPC) problems, conventional supervised machine learning methods (e.g., artificial neural networks and support vector machines) would have some difficulties. For instance, a training dataset containing both in-control and out-of-control process observations is required by a supervised machine learning method, but it is rarely available in SPC applications. Furthermore, many machine learning methods work like black boxes. It is often difficult to interpret their learning mechanisms and the resulting decision rules in the context of an application. In the SPC literature, there have been some existing discussions on how to handle the lack of out-of-control observations in the training data, using the one-class classification, artificial contrast, real-time contrast, and some other novel ideas. However, these approaches have their own limitations to handle SPC problems. In this paper, we extend the self-starting process monitoring idea that has been employed widely in modern SPC research to a general learning framework for monitoring processes with serially correlated data. Under the new framework, process characteristics to learn are well specified in advance, and process learning is sequential in the sense that the learned process characteristics keep being updated during process monitoring. The learned process characteristics are then incorporated into a control chart for detecting process distributional shift based on all available data by the current observation time. Numerical studies show that process monitoring based on the new learning framework is more reliable and effective than some representative existing machine learning SPC approaches.

Towards Double Robustness Under the Cox Marginal Structural Model

Xu, Ronghui (Lily)
Department of Mathematics,
University of California – San Diego

The Cox Marginal Structural Model (MSM) has been widely used to draw causal inference from survival data of observational studies. The typical estimation approach under the MSM is inverse probability weighting (IPW), which is known to be inefficient and inconsistent when the IP weights are not correctly estimated. Due to the non-collapsibility of the Cox regression model, double robustness (DR) is not a straightforward and possibly an ill-posed problem. We explore the extent to which, both theoretically and empirically, the augmented IPW (AIPW) may achieve DR. In the process we also carefully analyze different ways of augmentation and their implications.

Nonstationary Spatial Process Modeling using Dimension Expansion

Yang, Hou-Cheng
U.S. Food and Drug Administration

It is increasingly understood that the assumption of stationarity is unrealistic for many spatial processes. In this article, we combine dimension expansion with a spectral method to model big non-stationary spatial fields in a computationally efficient manner. Specifically, we use Mejía and Rodríguez-Iturbe's (1974) spectral simulation approach to simulate a spatial process with a covariogram at locations that have an expanded dimension. We introduce Bayesian hierarchical modeling to dimension expansion, which originally has only been modeled using a method of moments approach. We consider a novel scheme to re-weight levels in a Bayesian spatial hierarchical model that allows one to use non-stationary spectral simulation within a collapsed Gibbs sampler. Our method is both full rank and non-stationary, and can be applied to big spatial data because it does not involve storing and inverting large covariance matrices. We demonstrate the wide applicability of our approach through simulation studies, and an application using ozone data obtained from the National Aeronautics and Space Administration (NASA).

Sensor-based Modeling and Optimization of Additive Manufacturing

Yang, Hui
Department of Industrial and Manufacturing Engineering
Pennsylvania State University

Additive manufacturing (AM) provides a greater level of flexibility to produce a 3D part with complex geometries directly from the design. However, the widespread application of AM is currently hampered by technical challenges in process repeatability and quality control. To enhance the in-process information visibility, advanced sensing is increasingly invested for real-time AM process monitoring. The proliferation of in-situ sensing data calls for the development of analytical methods for the extraction of features sensitive to layerwise defects, and the

exploitation of pertinent knowledge about defects for in-process quality control of AM builds. As a result, there are increasing interests and rapid development of sensor-based models for the characterization and estimation of layerwise defects in the past few years. However, very little has been done to go from sensor-based modeling of defects to the suggestion of in-situ corrective actions for quality control of AM builds. In this talk, we present a new sequential decision-making framework for in-situ control of AM processes through the constrained Markov decision process (CMDP), which jointly considers the conflicting objectives of both total cost (i.e., energy or time) and build quality. Experimental results show that the CMDP formulation provides an effective policy for executing corrective actions to repair and counteract incipient defects in AM before completion of the build.

Online Sequential Monitoring of Spatio-Temporal Disease Incidence Rates

Yang, Kai
Department of Biostatistics
University of Florida

Online surveillance of infectious diseases, cancers and other deadly diseases is critically important for public health and safety of our society. Incidence data of such diseases are often collected spatially from different clinics and hospitals through regional, national and global disease reporting systems. In these systems, new batches of data keep being collected over time, and conventional statistical process control charts are usually included for decision making regarding whether there is a disease outbreak at the current time point after new data are collected. However, these control charts require many assumptions on the observed data, including the ones that the observed data are independent at different locations and/or times, and they follow a parametric distribution when no disease outbreaks are present. These assumptions are rarely valid in practice, making the results from the conventional control charts unreliable. Motivated by an application to monitor the Florida influenza-like illness data, we develop a new sequential monitoring approach, which can accommodate seasonality, spatio-temporal data correlation, and nonparametric data distribution. It is shown that the new method is more reliable to use in practice than the commonly used conventional charts for sequential monitoring of disease incidence rates.

A generic online nonparametric monitoring and sampling strategy for high-dimensional heterogeneous processes

Ye, Honghan
University of Wisconsin, Madison

With the rapid advancement of in-process measurements and sensor technology driven by zero-defect manufacturing applications, high-dimensional heterogeneous processes that continuously collect distinct physical characteristics frequently appear in modern industries. Such large-

volume high-dimensional data place a heavy demand on data collection, transmission, and analysis in practice. Thus, practitioners often need to decide which informative data streams to observe given the resource constraints at each data acquisition time, which poses significant challenges for multivariate statistical process control and quality improvement. In this article, we propose a generic online nonparametric monitoring and sampling scheme to quickly detect mean shifts occurring in heterogeneous processes when only partial observations are available at each acquisition time. Our innovative idea is to seamlessly integrate the Thompson sampling (TS) algorithm with a quantile-based nonparametric cumulative sum (CUSUM) procedure to construct local statistics of all data streams based on the partially observed data. Furthermore, we develop a global monitoring scheme of using the sum of top-$\boldsymbol{r}$ local statistics, which can quickly detect a wide range of possible mean shifts. Tailored to monitoring the heterogeneous data streams, the proposed method balances between exploration that searches unobserved data streams for possible mean shifts and exploitation that focuses on highly suspicious data streams for quick shift detection. Both simulations and a case study are comprehensively conducted to evaluate the performance and demonstrate the superiority of the proposed method.

A unified theory for covariate adjustment in the design and analysis of randomized clinical trials

Ye, Ting
Department of Statistics
University of Pennsylvania

Regulatory agencies highly encourage adjustments for baseline covariates at both the design and analysis stages of a clinical trial. At the design stage, covariate-adaptive randomization (CAR) is most widely applied as balance of treatment groups with respect to prognostic factors can enhance the credibility of the results (EMA, 2015); at the analysis stage, incorporating prognostic factors can result in a more efficient use of data to demonstrate and quantify the treatment effect (FDA, 2019). However, covariate adjustments pose two major challenges: (1) treatment assignment is no longer independent under CAR; (2) model-based approaches such as the Cox regression is unreliable when the model is mis-specified. In this talk, I will present a unified theory for how covariate adjustments at the two different stages affect the testing of treatment effect in survival analysis. Besides robustness and efficiency, I argue that universality – a valid testing procedure can be universally applicable to all commonly used randomization schemes – is another key aspect in evaluating analysis methods. Our theory addresses the analysis challenges arising from complex dependence in the treatment allocation and provides important guidance for the practice.

Nonparametric Denoising of Image Sequences

Yi, Fan
Department of Biostatistics
University of Florida

To monitor the Earth's surface, the satellite of the NASA Landsat program provides us image sequences of any region on the Earth constantly over time. These image sequences give us a unique resource to study the Earth surface, changes of the Earth resource over time, and their implications in agriculture, geology, forestry, and more. Besides natural sciences, image sequences are also commonly used in functional magnetic resonance imaging (fMRI) of medical studies for understanding the functioning of brains and other organs. In practice, observed images almost always contain noise and other contaminations. For a reliable subsequent image analysis, it is important to remove such contaminations in advance. This study focuses on image sequence denoising, which has not been well discussed in the literature yet. To this end, an edge-preserving image denoising procedure is suggested. The suggested method is based on a jump-preserving local smoothing procedure, in which the bandwidths are chosen such that the possible spatio-temporal correlation in the observed image intensities is accommodated properly. Both theoretical arguments and numerical studies show that it works well in various cases considered.

Building end-to-end sustainment models for weapon systems

Yi, Han
Institute for Defense Analyses

Improving readiness of weapon systems has been a persistent challenge for the Department of Defense (DoD). A key piece of this complex task is sustainment of large numbers of spares, personnel, operations, and maintenance that are needed for weapon systems to accomplish missions around the world. Currently, there is no DoD-wide tool to forecast the impact of specific sustainment investments on readiness outcomes. To this end, IDA has been working with military and civilian sponsors to develop end-to-end simulation models that fully incorporate the interconnected pieces of sustainment.

In this conference presentation, we describe the complexities of properly emulating sustainment concepts and argue for the urgency of end-to-end modeling efforts in investigating the strategic levers that drive readiness across the DoD. We discuss the challenges in aggregating and interpreting sustainment data to generate component-level metrics and how we employ statistical best practices to overcome these challenges. Next, we showcase notional examples of specific investments. Finally, we illustrate how examining the combined effects of multiple investments is pivotal in guiding senior decision-makers to make better investments to improve readiness.

An effective method for online disease risk monitoring

You, Lu
Health Informatics Institute
University of South Florida

Some deadly diseases can be treated or even prevented if they or some of their symptoms are detected early. Disease early detection and prevention is thus important for our health improvement. In this paper, we suggest a novel and effective new method for disease early detection. By this method, a patient's risk to the disease is first quantified at each time point by survival data analysis of a training dataset that contains patients' survival information and longitudinally observed disease predictors (e.g., disease risk factors and other covariates). To improve the effectiveness of the proposed method, variable selection is used in the survival analysis to keep only important disease predictors in disease risk quantification. Then, the longitudinal pattern of the quantified risk is monitored sequentially over time by a nonparametric control chart. A signal will be given by the chart once the cumulative difference between the risk pattern of the patient under monitoring and the risk pattern of a typical person without the disease in concern exceeds a control limit.

Joint work with Peihua Qiu.

Approximate Tolerance Regions for Simultaneous Autoregressive Models

Young, Derek
Department of Statistics
University of Kentucky

Spatial dependence in data can be characterized through any of numerous models that fall under the broad class of spatial regression models. Prediction from such estimated spatial regression models is often crucial for the application at hand. One possibility is to provide bounds that capture a specified proportion (or content) of such spatially-dependent data with a given level of confidence, which is the definition of a statistical tolerance interval. This talk provides, perhaps, one of the first attempts at developing statistical tolerance regions for any type of spatial regression model. We develop pointwise tolerance intervals for the simultaneous autoregressive model. A coverage study will highlight the general performance of the proposed approach. These tolerance intervals, with interpretations, will be applied to a dataset involving neighborhood-level mental health in the city of Seattle.

A Brief Tutorial on the mixtools and tolerance Packages for R

Young, Derek
Department of Statistics
University of Kentucky

The mixtools package includes visualization and estimation tools for various parametric and semiparametric mixture models. The tolerance package provides functions for calculating and visualizing statistical tolerance regions. These two packages are both over a decade old, being first released in 2006 and 2009, respectively. Evidence from industry sources and publications show that both packages are used extensively in areas like engineering, pharmaceuticals, medical device development, and quality control. Even though both packages are updated with some regularity, natural advancements in R and, more generally, software development tools, have underscored the need for more substantial modernization to these packages. In this tutorial, the capabilities of both packages will be presented while highlighting recently-added functions and modernizations to their respective visualization tools. Real examples using each package will be analyzed.

Detecting special-cause variation 'events' from process data signatures

Young, Timothy M.
Department of Forestry, Wildlife and Fisheries
University of Tennessee

The ability to detect special-cause variation of incoming feedstocks from advanced sensor technology is invaluable to manufacturers. Many on-line sensors produce data signatures that require further off-line statistical processing for interpretation by operational personal. However, early detection of changes in variation in incoming feedstocks may be imperative to promote 'early-stage' preventive measures. A method is proposed in this applied study for developing 'control bands' to quantify the variation of 'data signatures' in the context of statistical process control (SPC). Control bands based on pointwise prediction intervals constructed from the 'Bonferroni Inequality' and Bayesian smoothing splines are developed. Applications using the control band method for data signatures from near infrared (NIR) spectroscopy scans of industrial fibers of Switchgrass (Panicum virgatum) used for biofuels production, Loblolly Pine (Pinus taeda) fibers for medium density fiberboard production, and formaldehyde (HCHO) emissions from particleboard were used. Simulations curves (k) of k = 100, k = 1000, and k = 10000 indicate that the Bonferroni method for detecting special-cause variation is closer aligned with the Shewhart definition of control limits when the pdfs are Gaussian or lognormal.

Power-Enhanced Simultaneous Test of High-Dimensional Mean Vectors and Covariance Matrices

Yu, Xiufan
Department of Applied and Computational Mathematics and Statistics
University of Notre Dame

Power-enhanced tests with high-dimensional data have received growing attention in theoretical and applied statistics in recent years. Various tests possess their respective high-power regions. In practice, we may lack prior knowledge about the alternatives when testing for a problem of interest. There is a critical need of developing powerful testing procedures against more general alternatives. In this talk, we will focus on power enhancement in jointly testing two-sample mean vectors and covariance matrices of high-dimensional data. Existing works mainly focus on testing mean vectors or covariance structures instead of testing both aspects together. To address this challenge, we present a new power-enhanced simultaneous test. The proposed test is powerful in detecting either mean differences or covariance differences under either sparse or dense alternatives. Theoretical analyses prove the accurate asymptotic size and consistent asymptotic power of the proposed test against more general alternatives, and simulation studies demonstrate the finite-sample performance. Moreover, we apply the proposed test in a real application to find differentially expressed gene-sets in cancer studies, which confirms prodigious performance of the proposed test with the support of biological evidence.

Min-Max Optimal Design of Two-Armed Trials with Side Information

Zhang, Qiong
School of Mathematical and Statistical Sciences
Clemson University

Significant evidence has become available that emphasizes the importance of personalization in medicine. In this work, we study the optimal design of two-armed clinical trials to maximize statistical accuracy where patient covariates are incorporated into treatment effect. We take a robust optimization approach and minimize (over design) the maximum (over covariates) variance of treatment effect. The resulting bi-level mixed integer nonlinear program is notably challenging to solve. We test our proposed algorithms with synthetic and real-world data sets and compare it with standard (re-)randomization methods.

Co-author: Amin Khademi, Yongjia Song

A Novel Robust Coupled CP Decomposition Method Based on Alternating Direction Method of Multipliers (ADMM)

Zhao, Meng
University of Florida

The coupled CANDECOMP/PARAFAC (CP) decomposition method is a powerful tool to extract a common hidden structure from heterogeneous tensors simultaneously. It has reached a crucial status in the areas of computer vision, signal processing, and neuroscience. However, to the best of our knowledge, most existing coupled CP decomposition methods are sensitive to outliers and lack robustness. This has motivated us to propose a novel robust coupled CP decomposition (RCCPD) method based on ADMM for simultaneously decomposing a pair of tensors with the presence of outliers. We compare our approach with the classical CP decomposition method and the coupled matrix and tensor factorization (CMTF) method. Experiments on both synthetical and real-world datasets demonstrate that the proposed RCCPD method can effectively handle outliers and jointly decompose the coupled tensors more accurately than compared approaches.

Notes

Notes

Notes

Notes

ASA®

MEMBERSHIP IS ONLY $25!

Student Resources

ASA student membership includes subscriptions to ***Amstat News*** and ***Significance*** magazines, an online subscription to ***CHANCE*** magazine, and access to the ASA's premier journals. Membership also includes networking within the ***ASA Community***; ASA Member News—the ASA's monthly e-newsletter; and information about internships, resources, and career opportunities.

STAT*tr@k*

This website offers tips for applying for a job, being a successful graduate student, and making the transition from coursework to research. The site also provides links to and information about career and mentorship sites, upcoming conferences, and awards and competitions. New articles appear monthly.

JobWeb

A targeted job database and résumé-posting service that helps you take advantage of valuable career opportunities.

JSM Career Service

The Joint Statistical Meetings offers the Career Service, a full-service recruiting facility for employers and job seekers. Each year, hundreds of applicants look to the service to begin or further their careers in statistics.

Internships

Every December, a list of internship opportunities for students is published in the ASA's membership magazine, *Amstat News*. The list also is posted on STAT*tr@k* and in the education area of the ASA's website.

Awards and Scholarships

The ASA's awards program recognizes student's contributions to statistics and offers paper competitions, travel awards, grants, and scholarships.

ASA Student Chapters

Student chapters encourage students to continue studying statistics and provide career information in the statistical sciences. Start or join a chapter and get free T-shirts with your chapter logo and funding to hold your own social events.

Like us on **Facebook** or follow us on **Twitter!**