International Indian Statistical Association
Annual Conference
Colorado School of Mines
Golden, Colorado
1st - 4th June, 2023

COLORADO SCHOOL OF MINES | Applied Mathematics and Statistics

KOREAN INTERNATIONAL STATISTICAL SOCIETY

Chair: Anindya ROY, University of Maryland, Baltimore County
Co-chair: Hiya BANERJEE, Eli Lilly
Co-chair: Rajatshi GUHANIYOGI, Teaxs A & M University

Scientific Programme Committee

AvantiATHREYA
Soutir BANDYOPADHYAY
Nairanjana DASGUPTA
Tanujit DEY
Bharani DHARAN
Yehenew KIFLE
Joyee GHOSH
Aditya GUNTUBOYINA
Yi HUANG
Kaushik JANA
Po-Ling LOH
Tucker MCELROY
Sai POPURI
Suchitrita SENGUPTA
Sandip SINHARAY

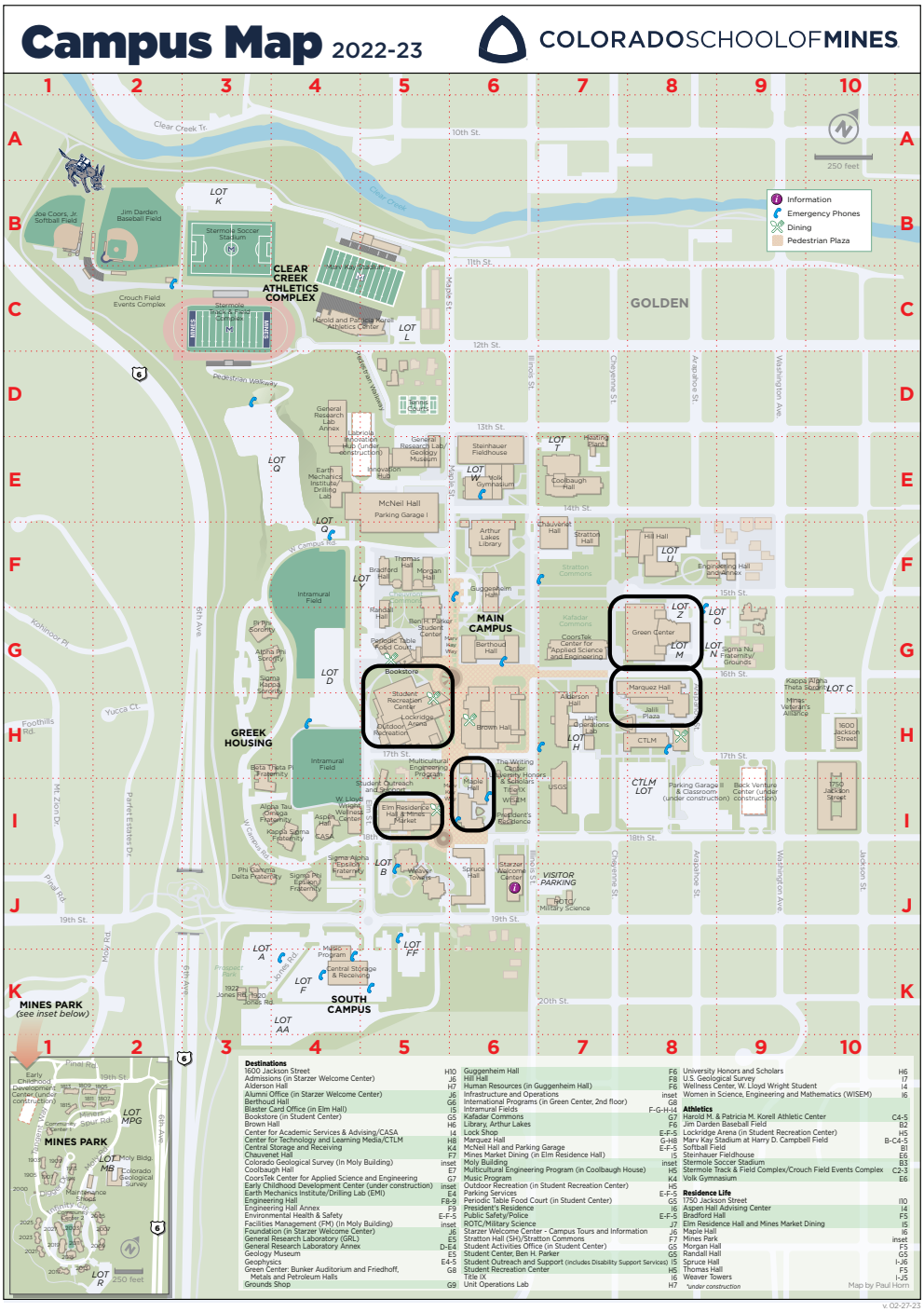
Local Organising Committee
Convenor: Soutir BANDYOPADHYAY, Colorado School of Mines
Dorit HAMMERLING
Andee KAPLAN
Rick ZHOU
William KLEIBER
Debashis GHOSH
Tusharkanti GHOSH

International Organizing Committee:
Snigdhansu CHATTERJEE, University of Minnesota
Subrata KUNDU, George Washington University
Subhashis GHOSHAL, North Carolina State University
Abhyuday MANDAL, University of Georgia
Hiya BANERJEE, Eli Lilly
Sanjay CHAUDHURI, University of Nebraska-Lincoln
Saonli BASU, University Minnesota
Debashis MONDAL, Washington University of St. Louis

Student and Staff Volunteers:
Sweta RAI, Maggie BAILEY, Prasoon GARG, Souvik BERA, Will DANIELS, Daniel Callan RICHARDS-DINGER, Olga KHALIUKOVA, Meng JIA.

Sponsorships and Endorsements:


Campus Map


Campus Map 2022-23
COLORADOSCHOOLOFMINES
Information
Emergency Phones
Dining
Pedestrian Plaza
250 feet
CLEAR CREEK ATHLETICS COMPLEX
GOLDEN
MAIN CAMPUS
GREEK HOUSING
SOUTH CAMPUS
MINES PARK
(see inset below)
VISITOR PARKING
Destinations
1600 Jackson Street H10
Admissions (in Starzer Welcome Center) J6
Alderson Hall H7
Alumni Office (in Starzer Welcome Center) J6
Berthoud Hall G6
Blaster Card Office (in Elm Hall) I5
Bookstore (in Student Center) G5
Brown Hall H6
Center for Academic Services & Advising/CASA I4
Center for Technology and Learning Media/CTLM H8
Central Storage and Receiving K4
Chauvenet Hall E7
Colorado Geological Survey (in Moly Building) inset
Coolbaugh Hall E7
CoorsTek Center for Applied Science and Engineering G7
Early Childhood Development Center (under construction) inset
Earth Mechanics Institute/Drilling Lab (EMI) E4
Engineering Hall F8-9
Engineering Hall Annex F9
Environmental Health & Safety E-F-5
Facilities Management (FM) (in Moly Building) inset
Foundation (in Starzer Welcome Center) J6
General Research Laboratory (GRL) E5
General Research Laboratory Annex D-E4
Geology Museum E5
Geophysics E4-5
Green Center: Bunker Auditorium and Friedhoff, Metals and Petroleum Halls G8
Grounds Shop G9
Guggenheim Hall F6
Hill Hall F8
Human Resources (in Guggenheim Hall) F6
Infrastructure and Operations inset
International Programs (in Green Center, 2nd floor) G8
Intramural Fields F-G-H-I4
Kafadar Commons G7
Library, Arthur Lakes F6
Lock Shop E-F-5
Marquez Hall G-H8
McNeil Hall and Parking Garage E-F-5
Mines Market Dining (in Elm Residence Hall) I5
Moly Building inset
Multicultural Engineering Program (in Coolbaugh House) H5
Music Program K4
Outdoor Recreation (in Student Recreation Center) H5
Parking Services E-F-5
Periodic Table Food Court (in Student Center) G5
President's Residence I6
Public Safety/Police E-F-5
ROTC/Military Science J7
Starzer Welcome Center - Campus Tours and Information J6
Stratton Hall (SH)/Stratton Commons F7
Student Activities Office (in Student Center) G5
Student Center, Ben H. Parker G5
Student Outreach and Support (includes Disability Support Services) I5
Student Recreation Center H5
Title IX I6
Unit Operations Lab H7
University Honors and Scholars H6
U.S. Geological Survey I7
Wellness Center, W. Lloyd Wright Student I4
Women in Science, Engineering and Mathematics (WISEM) I6
Athletics
Harold M. & Patricia M. Korell Athletic Center C4-5
Jim Darden Baseball Field B2
Lockridge Arena (in Student Recreation Center) H5
Marv Kay Stadium at Harry D. Campbell Field B-C4-5
Softball Field B1
Steinhauer Fieldhouse E6
Stermole Soccer Stadium B3
Stermole Track & Field Complex/Crouch Field Events Complex C2-3
Volk Gymnasium E6
Residence Life
1750 Jackson Street I10
Aspen Hall Advising Center I4
Bradford Hall F5
Elm Residence Hall and Mines Market Dining I5
Maple Hall I6
Mines Park inset
Morgan Hall F5
Randall Hall G5
Spruce Hall I-J6
Thomas Hall F5
Weaver Towers I-J5
*under construction
Map by Paul Horn
v. 02-27-23

Information

Dorm and Venue Addresses:

Dorm address:
Maple Hall,
Colorado School of Mines
1733 Maple Street,
Golden, CO 80401

Maple Hall Front Desk:
Phone: 303-273-3984
Monday – Friday: 8:00AM – 8:00PM
Saturday – Sunday: 10:00AM – 4:00PM; 6:00PM – 8:00PM

Mines Market (For breakfast):
Mines Market, Elm Hall
1795 Elm St.,
Golden, CO 80401

Ben H. Parker Student Center (Student Center):
1200 16th Street,
Golden, CO 80401

Green Center (Friedhoff Halls):
924 16th Street,
Golden, CO 80401

Marquez Hall:
1600 Arapahoe St,
Golden, CO 80401

Food Coupon Color Codes:

- Red, Pink, Orange: Lunch coupons for June 1, 2, and 3 respectively.
- Green: Banquet dinner coupon (June 2).
- White: Wine coupons for banquet (June 2).

Internet Access Options:

- Mines-Guest (Open Access)
- Eduroam

From The President, IISA

The International Indian Statistical Association (IISA) and the Applied Mathematics and Statistics Department of the Colorado School of Mines are going to organize the 2023 IISA Conference on June 1-4 in Golden, Colorado, a suburb of the thriving metropolis of Denver, at the foothills of the beautiful Rocky Mountains. On behalf of IISA, I am writing to welcome you to the conference. IISA is a U.S.-based non-profit organization that aims to promote education, research, and application of statistics, probability, and data science and foster the exchange of information and scholarly activities worldwide, with a particular emphasis on the Indian subcontinent. IISA membership is open to individuals of all nationalities, races, and genders and is not restricted to people of Indian descent. IISA is committed to serving the needs of young statisticians and data scientists and promoting diversity. IISA encourages cooperative efforts among members in education, research, industry, and business.

As the world was returning to life after the COVID-19 pandemic, IISA successfully organized its 2022 annual conference in Bengaluru, India, where more than four hundred delegates participated in person. Pursuing the IISA tradition of alternating between India and North America, IISA 2023 will take place in the U.S. It will provide a unique platform for conference participants to exchange ideas on recent progress in statistical theory, methodology, and applications, and integration with data science. To promote research by students, IISA 2023 will organize student paper and poster competitions and award the winners with prizes. The rich program consists of a Bahadur Memorial Lecture, four other plenary talks, twelve special invited talks, about two hundred invited talks, several contributed talks, three panel discussion sessions, three short courses, and the student paper and poster sessions. In addition, a session devoted to the collaboration among statisticians of African and Indian descent and a panel discussion session on women's issues sponsored by the ASA Committee on Women in Statistics will help promote diversity.

The IISA 2023 conference is a huge endeavor, and we acknowledge the work and efforts of the local and scientific committee members, session organizers, speakers, student paper and poster competition judges, and faculty, staff, and student volunteers at the local host, The Colorado School of Mines. The scientific program committee, led by Professor Anindya Roy, created an exciting program. Professor Soutir Bandyopadhyay is in charge of the local organizing committee, which has been working tirelessly to make everything perfect. Several members of the IISA Executive Committee and IISA volunteers spent numerous hours making the conference successful. Finally, participation makes a conference successful. We are looking forward to welcoming all IISA 2023 participants to the Colorado School of Mines Campus on June 1-4, 2023.

Subhashis Ghoshal
President, International Indian Statistical Association
Goodnight Distinguished Professor of Statistics, North Carolina State University

AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

May 17, 2023

Dear Friends,

What brings us together is our shared commitment to advancing the statistical science. What makes us a community is our shared commitment to building cross-cultural alliances. The ASA and IISA share a rich history of significant collaboration, so it is my honor to offer greetings and best wishes from the ASA Board of Directors and the broader membership as you come together at the IISA 2023 Conference at the Colorado School of Mines in Golden.

I note with pleasure you have invited sessions from ICSA and KISS and are featuring the contributions of African statisticians. This intentional inclusiveness matches the ASA's goals and my goals as ASA President as well.

The excellent program will energize the participants creating a productive and memorable conference. I congratulate you on offering a program that celebrates the importance and diversity of our science.

Sincerely,

Dionne Price, 2023 President, ASA

Contents

Program Overview

Program

Thursday June 1

Registration

Venue: SC Lobby

Time : **8:00 - 18:00**

Conference Inaugaration

Venue: SC Ball A

Time : **8:30 - 8:50**

- **Terry HOGUE**, Dean, Earth and Society Programs, Colorado School of Mines

01.M1.I1 Privacy-preserving data integration for statistical inference

Venue: SC Ball B

Chair and Organizer : Srijan SENGUPTA, North Carolina State University

9:00 **Piecewise and distributed learning using federated methods: statistical considerations and operating characteristics** [Abstract 18]

Anjishnu BANERJEE, *Medical College of Wisconsin*

9:25 **Bayesian methods for vaccine safety surveillance using federated data sources** [Abstract 40]

Fan BU, *UCLA*

9:50 **Statistical Optimality of Federated Learning Beyond Stationary Points** [Abstract 237]

Jiaming XU, *Duke University*
Lili SU, *Northeastern University*
Pengkun YANG, *Tsinghua University*

01.M1.I2 Causal Inference

Venue: SC Ball C

Chair and Organizer : Adityanand GUNTUBOYINA, University of California Berkeley

9:00 **Balancing Weights for Causal Inference in Observational Factorial Studies** [Abstract 241]

Ruoqi YU, *University of California, Davis*
Peng DING, *University of California, Berkeley*

9:25 **Covariate-adaptive randomization inference in matched designs** [Abstract 183]

Sam PIMENTEL, *UC Berkeley*

9:50 **A new central limit theorem for the augmented IPW estimator: variance inflation, cross-fit covariance and beyond** [Abstract 169]

Rajarshi MUKHERJEE, *Harvard T.H. Chan School of Public Health*
Kuanhao JIANG, *Harvard University*
Subhabrata SEN, *Harvard University*
Pragya SUR, *Harvard University*

01.M1.I3 Recent advances in survey sampling

Venue: SC Ball E

Chair and Organizer : Jae-kwang KIM, Iowa State University

9:00 **Combining Data Sources to Produce Nationally Representative Estimates of Hospital Encounter Characteristics** [Abstract 37]

Jay BREIDT, *NORC at the University of Chicago*
Dean RESNICK, *NORC at the University of Chicago*

Geoffrey JACKSON, *National Center for Health Statistics*
Donielle WHITE, *National Center for Health Statistics*

9:25 **On a modified deep neural network based mass imputation for data integration** [**Abstract 60**]

Sixia CHEN, *University of Oklahoma Health Sciences Center*

9:50 **Semiparametric adaptive estimation under informative sampling** [**Abstract 167**]

Kosuke MORIKAWA, *Osaka University*
Jae Kwang KIM, *Iowa State University*
Yoshikazu TERADA, *Osaka University*

01.M1.I4 Advances in Extreme Value Analysis

Venue: MZ 122

Chair : Trevor HARRIS, Texas A&M University
Organizer : Reetam MAJUMDER, NC State University

9:00 **Accounting for the spatial structure of weather systems in detected changes in precipitation extremes** [**Abstract 244**]

Likun ZHANG, *University of Missouri*
Mark RISSER, *Lawrence Berkeley National Laboratory*
Edward MOLTER, *University of California, Berkeley*
Michael WEHNER, *Lawrence Berkeley National Laboratory*

9:25 **A Deep Learning Synthetic Likelihood Approximation of a Non-stationary Spatial Model for Extreme Streamflow Forecasting** [**Abstract 153**]

Reetam MAJUMDER, *NC State University*
Brian J. REICH, *NC State University*

9:50 **Characterizing Asymptotic Dependence between a Satellite Precipitation Product and Station Data in the Northern US Rocky Mountains via the Tail Dependence Regression Framework with a Gibbs Posterior Inference Approach** [**Abstract 195**]

Brook RUSSELL, *Clemson University School of Mathematical and Statistical Sciences*
Yiren DING, *Clemson University School of Mathematical and Statistical Sciences*
Whitney HUANG, *Clemson University School of Mathematical and Statistical Sciences*
Jamie DYER, *Department of Geosciences, Mississippi State University*

01.M1.I5 Topics on Causal Inference and Mediation Analysis

Venue: MZ 226

Chair : Xiaofeng WANG, Cleveland Clinic
Organizer : Tanujit DEY, Harvard Medical School

9:00 **Semiparametric Regression Models for Causal Mediation Analysis with Longitudinal Data** [**Abstract 4**]

Jeffrey ALBERT, *Case Western Reserve University*
Tanujit DEY, *Harvard Medical School and Brigham and Women's Hospital*
Hongxu ZHU, *Case Western Reserve University*
Jiayang SUN, *Ceorge Mason University*

9:25 **Combining multiple sources of information to estimate hearing loss prevalence in the United States at the county level by gender, age, and race/ethnicity using small area estimation models** [**Abstract 84**]

Carolina FRANCO, *NORC at the University of Chicago*
David REIN,

9:50 **Estimating heterogeneous treatment effects on binary outcomes with noncompliance using Bayesian additive regression trees** [Abstract 72]

Sameer DESHPANDE, *University of Wisconsin–Madison*
Jared FISHER, *Brigham Young University*
David PUELZ, *University of Texas at Austin*

01.M1.I6 Bayesian methods for multimodal data integration

Venue: MZ 235

Chair : Miheer DEWASKAR, Duke University
Organizer : Himel MALLICK, Cornell University

9:00 **Bayesian cooperative learning with BART** [Abstract 22]

Piyali BASAK, *Merck & Co.*
Himel MALLICK, *Cornell University*
Erina PAUL, *Merck & Co.*

9:25 **Joint Additive Factor Regression for Multi-Omics Data Integration** [Abstract 8]

Niccolo ANCESCHI, *Duke University*
Federico FERRARI, *Merck & Co., Inc.*
Himel MALLICK, *Cornell University*
David DUNSON, *Duke University*

9:50 **Global-local Shrinkage Priors for Multimodal Data Integration** [Abstract 159]

Omar MELIKECHI, *Duke University*

Student Paper Competition 1 Application of Statistics and Data Sciences

Venue: SC Ball D

9:00 **Incorporating Interventions to an Extended SEIRD Model with Vaccination: Application to COVID-19 in Qatar** [Abstract 7]

Elizabeth B AMONA, *Virginia Commonwealth University*

9:15 **Robust probabilistic inference via a constrained transport metric** [Abstract 48]

Abhisek CHAKRABORTY, *Texas A&M University*

9:30 **Achieving Privacy-Utility Balance in Time Series Release Mechanisms Using Multiple Imputation and Stochastic Filtering** [Abstract 116]

Gaurab HORE, *University of Maryland, Baltimore County*

9:45 **Fast Parameter Estimation of GEV using Neural Networks** [Abstract 188]

Sweta RAI, *Colorado School of Mines*

10:00 **Probabilistic Inverse Model: An Application in Hydrology** [Abstract 213]

Somya SHARMA, *University of Minnesota - Twin Cities*

Coffee 10:20 - 10:50 Venue: Avery Room at the Student Center

Special Session 1 Special Session on Mathematics and Statistics Education: Learning Mathematics and Statistics

Venue: SC Ball A

Chair : Ansu CHATTERJEE, University of Minnesota

- **Bootstrapping for Learning Statistics** [Abstract 113]

  Tim HESTERBERG, *Instacart*

- **Po-Shen LOH**, Carnegie Mellon University

01.M2.I7 Statistical Anecdotes in Neuroscience, Rare Disease, and Cancer Drug Development

Venue: SC Ball B

Chair and Organizer : Arnab MAITY, Pfizer

10:50 **Opportunities and challenges in neuroscience clinical trials** [Abstract 228]

Ling WANG, *Alkermes Inc*

11:15 **Case Weighted Adaptive Power Priors for Hybrid External Control Arms** [Abstract 141]

Evan KWIATKOWSKI, *University of Texas MD Anderson Cancer Center*

11:40 **The Kumaraswamy distribution as a failure model under various loss functions and based on progressive censored data** [Abstract 111]

Amal HELU, *The University of Jordan*

01.M2.I8 Addressing challenges of nonstandard time-to-event endpoints in clinical trials

Venue: SC Ball C

Chair and Organizer : Amarjot KAUR, Merck Research Labs

10:50 **How to compare two curves: the easiest question in survival analysis?** [Abstract 52]

Rick CHAPPELL, *University of Wisconsin*
Mitchell PAUKNER, *University of Wisconsin*

11:15 **Means as outcomes: Improving interpretation when hazards are non-proportional.** [Abstract 181]

Mitchell PAUKNER, *Northwestern University*

11:40 **Evaluation of Log-rank, RMST and MaxCombo in Immuno-Oncology(IO) trials – A Retrospective Analysis in Patients Treated with Anti-PD1/PD-L1 Agents across Solid Tumors** [Abstract 239]

Jiabu YE, *Merck Research Labs*

01.M2.I9 Recent advances in network data inference

Venue: SC Ball E

Chair and Organizer : Srijan SENGUPTA, North Carolina State University

10:50 **Discovering underlying dynamics in time series of networks** [Abstract 12]

Avanti ATHREYA, *Johns Hopkins University*
Zachary LUBBERTS, *Johns Hopkins University*
Youngser PARK, *Johns Hopkins University*
Carey PREIBE, *Johns Hopkins University*

11:15 **Model selection for network data based on spectral information** [Abstract 221]

Jonathan STEWART, *Florida State University*
Jairo PEÑA, *Florida State University*

11:40 Conformal Prediction for Network-Assisted Regression [Abstract 149]

Robert LUNDE, *Washington University in St. Louis*
Elizaveta LEVINA, *University of Michigan*
Ji ZHU, *University of Michigan*

01.M2.I10 Methods and computing for large spatial data

Venue: MZ 122
Chair and Organizer : Reetam MAJUMDER, NC State University

10:50 Variational sparse inverse Cholesky approximation for latent Gaussian processes via double Kullback-Leibler minimization [Abstract 44]

Jian CAO, *Texas A&M University*
Myeongjong KANG, *Texas A&M*
Felix JIMENEZ, *Texas A&M*
Matthias KATZFUSS, *Texas A&M*

11:15 Inverses of Matern Covariances on Grids [Abstract 105]

Joseph GUINNESS, *Cornell University*

11:40 Multi-model Ensemble Analysis with Neural Network Gaussian Processes [Abstract 109]

Trevor HARRIS, *Texas A&M University*

01.M2.I11 Recent Advances in Network Analysis and Spatial Profiling

Venue: MZ 235
Chair : Rajarshi GUHANIYOGI, Texas A&M University
Organizer : Sharmistha GUHA, Texas A&M University

10:50 Multilayered Network Models for Security: Enhancing System Security Engineering with Orchestration [Abstract 234]

Adam D. WILLIAMS, *Principal R&D Systems Engineer, Sandia Laboratories*

11:15 A Bayesian Approach to Network Classification [Abstract 101]

Sharmistha GUHA, *Texas A&M University*
Abel RODRIGUEZ, *University of Washington*

Student Paper Competition 2 Probability and Theory of Statistics and Data Sciences

Venue: SC Ball D

10:50 MARS via LASSO [Abstract 130]

Dohyeong KI, *UC Berkeley*

11:05 Optimal Bayesian Inference for High-dimensional Linear Regression Based on Sparse Projection-posterior [Abstract 176]

Samhita PAL, *North Carolina State University*

11:20 Global-Local Shrinkage Priors for Asymptotic Point and Interval Estimation of Normal Means under Sparsity [Abstract 187]

Zikun QIN, *University of Florida*

11:35 Limit theorems for semi-discrete optimal transport maps [Abstract 196]

Ritwik SADHU, *Department of Statistics and Data Science, Cornell.University*

11:50 Coverage of Credible Intervals in Bayesian Multivariate Isotonic Regression [Abstract 227]

Kang WANG, *North Carolina State University*

Lunch 12:10 - 13:30 Venue: Avery Room at the Student Center

Short Course 1 Statistical Inference of Network Data: Past, Present, and Future *Venue:* SC Ball E

Time : **13:30 - 17:00**

- **Srijan SENGUPTA**, North Carolina State University

Special Invited Session 1 Catherine Calder, Mary Meyer *Venue:* SC Ball A

Chair : Hiya BANERJEE, Eli Lilly

13:30 Statistical and Ethical Considerations in the Analysis of Mobile Phone Tracking Data [Abstract 43]

Catherine CALDER, *University of Texas at Austin*
Marcin JUREK, *University of Texas at Austin*
Corwin ZIGLER, *University of Texas at Austin*

14:10 Constrained Spline Density Estimation and Applications [Abstract 160]

Mary MEYER, *Colorado State University*
Hanxiao JING, *Colorado State University*
Xin CHEN,

01.A1.I12 Statistical Inference for multivariate and High dimensional data *Venue:* SC Ball B

Chair : Dan NORDMAN, Iowa State University
Organizer : Soumendranath LAHIRI, Washington university of St. Louis

13:30 A bootstrap test for MANOVA in high dimensions [Abstract 49]

Nilanjan CHAKRABORTY, *Washington University in Saint Louis*
Lyudmila SAKHANENKO, *Michigan State University*

13:55 Post-selection inference for regression with grouped responses [Abstract 98]

Karl GREGORY, *University of South Carolina*
Qinyan SHEN, *University of South Carolina*
Xianzheng HUANG, *University of South Carolina*

14:20 A Bootstrap based Goodness-of-fit test of covariance for multiple outcomes in Longitudinal Data [Abstract 90]

Dhrubajyoti GHOSH, *Duke University*
Sheng LUO, *Duke University*

01.A1.I13 New developments in high dimensional inferences for dependent data

Venue: SC Ball C

Chair and Organizer : Arkaprava ROY, University of Florida

13:30 Prediction Interval for high-dimensional regression with dependent errors [Abstract 125]

Sayar KARMAKAR, *University of Florida*

13:55 Scalable Nonparametric Bayesian Learning for Dynamic Velocity Fields [Abstract 99]

Aritra GUHA, *AT&T Data Science and AI Research*
Sunrit CHAKRABORTY, *University of Michigan*
Rayleigh LEI, *University of Washington*
XuanLong NGUYEN, *University of Michigan*

14:20 Sparse canonical correlation for integrative multi-omics explain pan-cancer and cancer-specific patterns of association [Abstract 79]

Diptavo DUTTA, *NIH/NCI*

01.A1.I14 Novel Statistical Approaches in Clinical Trials

Venue: MZ 122

Chair and Organizer : Bharani DHARAN, Novartis Pharmaceuticals

13:30 Challenges of estimating individual treatment effects from clinical data using machine learning [Abstract 148]

Ilya LIPKOVICH, *Eli Lilly and Company*
David SVENSSON, *AstraZeneca*
Bohdana RATITCH, *Bayer*
Alex DMITRIENKO, *Mediana*

13:55 Probability-of-Decision Designs to Accelerate Dose-Finding Trials [Abstract 246]

Tianjian ZHOU, *Colorado State University*

14:20 How to use causal inference with clinical trial data of CAR-T cell therapies [Abstract 150]

Wanying MA, *Novartis Pharmaceuticals Corporation*
Edward WALDRON, *Novartis Pharmaceuticals Corporation*
Julie JONES, *Novartis Pharma AG*

01.A1.I15 Statistical Challenges and Opportunities in the Biopharmaceutical Industry

Venue: MZ 226

Chair : Piyali BASAK, Merck & Co.
Organizer : Arinjita BHATTACHARYYA, Merck & Co., Inc.

13:30 Opportunities and Challenges in Using External Information in Drug Development [Abstract 128]

Amarjot KAUR, *Merck Research Labs*
Bhramori BANERJEE, *Merck*

13:55 Discussions and Challenges in Multi-Arm Multi-Stage (MAMS) Designs [Abstract 194]

Niladri ROY CHOWDHURY, *Bristol-Myers-Squibb*
Xue WU, *Penn State University*
Arun KUMAR, *Bristol-Myers-Squibb*

14:20 **Cumulative Logistic Ordinal Regression with Proportional Odds when the Missing Responses are Nonignorable – Application to Phase III Trial** [Abstract 152]

Arnab MAITY, *Pfizer*
Huaming TAN, *Pfizer*
Vivek PRADHAN, *Pfizer*
Soutir BANDYOPADHYAY, *Colorado School of Mines*

01.A1.I16 Precision medicine in cancer research

Venue: MZ 235

Chair : Anjishnu BANERJEE, Medical College of Wisconsin
Organizer : Tanujit DEY, Harvard Medical School

13:30 **Current methods for evaluating prediction model performance** [Abstract 126]

Michael KATTAN, *Department of Quantitative Health Sciences, Cleveland Clinic*

13:55 **Use of longitudinal serum markers as early predictors of treatment outcome in germ cell cancers** [Abstract 180]

Sujata PATIL, *Cleveland Clinic*

14:20 **Robust and replicable supervised and unsupervised learning methods for cancer precision medicine** [Abstract 190]

Naim RASHID, *Department of Biostatistics, Gillings School of Global Public Health, UNC-CH*

Short Break 14:50 - 15:00

Special Invited Session 2 Amita Manatunga, David Ohlssen

Venue: SC Ball A

Chair : Bharani DHARAN, Novartis Pharmaceuticals

15:00 **Development and evaluation of a computer-aided diagnosis system (CAD) in the absence of a gold standard.** [Abstract 156]

Amita MANATUNGA, *Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University*

15:40 **Solving Drug Development Challenges with Data Science: Combining Statistical Inference, Modeling, and Machine Learning for Effective Decision Making** [Abstract 174]

David OHLSSEN, *Novartis*

Panel Discussion 1 Let's Talk Issues

Venue: MZ 235

Organizer : Nairanjana DASGUPTA, Washington State University

Sponsor: ASA Committee on Women in Statistics

- **Elizabeth Juarez COLUNGA**, GRECC
- **Nairanjana DASGUPTA**, Washington State University
- **Sharmistha GUHA**, Texas A&M University
- **Hiya BANERJEE**, Eli Lilly

01.A2.I17 Methods for Highly Multivariate Nonstationary Spatial Processes

Venue: SC Ball B

Chair : Soutir BANDYOPADHYAY, Department of Applied Mathematics and Statistics, Colorado School of Mines

Organizer : Tucker MCELROY, US Census Bureau

15:00 Modeling massive highly-multivariate nonstationary spatial data with the basis graphical lasso [Abstract 134]

William KLEIBER, *University of Colorado Boulder*
Mitchell KROCK, *University of Colorado Boulder*
Dorit HAMMERLING, *Colorado School of Mines*
Stephen BECKER, *University of Colorado Boulder*

15:25 Regridding uncertainty for statistical downscaling of solar radiation. [Abstract 17]

Soutir BANDYOPADHYAY, *Department of Applied Mathematics and Statistics, Colorado School of Mines*

15:50 Development of a Nonstationary Spatio-Temporal Model for Water Vapor Flux Divergence [Abstract 123]

Mark KAISER, *Iowa State University*
Jonathan HOBBS, *NASA Jet Propulsion Laboratory, California Institute of Technology*

01.A2.I18 Data aggregation, non-probability surveys and applications in government statistics

Venue: SC Ball D

Chair : Jean OPSOMER, Westat

Organizer : Ansu CHATTERJEE, University of Minnesota

15:00 Statistical data integration using multilevel models [Abstract 82]

Andreea ERCIULESCU, *Westat*
Jean OPSOMER, *Westat*
Benjamin SCHNEIDER, *Westat*

15:25 Joint Point and Variance Estimation under a Hierarchical Bayesian model for Survey Count Data [Abstract 204]

Terrance SAVITSKY, *BLS*

15:50 Maximum likelihood estimation of response propensity to a nonprobability survey [Abstract 26]

Vladislav BERESOVSKY, *Bureau of Labor Statistics*
Terrance SAVITSKY, *BLS*
Matthew WILLIAMS, *RTI*
Julie GERSHUNSKAYA, *BLS*

01.A2.I19 Modern approaches in statistical neuroimaging

Venue: MZ 122

Chair and Organizer : Rajarshi GUHANIYOGI, Texas A&M University

15:00 Bayesian image analysis in Fourier space [Abstract 139]

John KORNAK, *University of California, San Francisco*
Karl YOUNG, *University of California, San Francisco (Retired)*
Eric FRIEDMAN, *University of California, Berkeley*

15:25 **Graph estimation in high dimensional time-series** [Abstract 191]

Arkaprava ROY, *University of Florida*
Anindya ROY, *UMBC*
Subhashis GHOSAL, *NCSU*

15:50 **A Bayesian Time-Varying Psychophysiological Interaction (PPI) Model** [Abstract 104]

Michele GUINDANI, *UCLA, Biostatistics*

01.A2.I20 Bayesian statistics in drug development and biomarker discovery

Venue: MZ 226

Chair and Organizer : Piyali BASAK, Merck & Co.

15:00 **Robustifying likelihoods by optimistically re-weighting data** [Abstract 73]

Miheer DEWASKAR, *Duke University*
Christopher TOSH, *Memorial Sloan Kettering Cancer Center*
Jeremias KNOBLAUCH, *UCL*
David DUNSON, *Duke University*

15:25 **Bayesian Survival Tree Ensembles with Submodel Shrinkage** [Abstract 147]

Antonio LINERO, *The University of Texas at Austin*

15:50 **A Bayesian approach for utilizing historical data in early drug development** [Abstract 87]

Nairita GHOSAL, *Merck & Co., Inc., Rahway, NJ, USA*

Memorial Session 1 In Memorium: Professor Dalho Kim

Venue: SC Ball C

Chair and Organizer : Gyuhyeong GOH, Kansas State University **Sponsor:** Korean International Statistical Association

15:00 **Interrelationship between Divergence Measures** [Abstract 92]

Malay GHOSH, *University of Florida*
Partha SARKAR, *University Of Florida*

15:25 **Bayesian framework for image analysis in aging studies** [Abstract 133]

Namhee KIM, *Rush University Medical Center*
Namhee KIM, *Rush University Medical Center*

15:50 **Bayesian Predictive Inference for Small Areas Using a Non-Probability Sample with Supplemental Information** [Abstract 170]

Balgobin NANDRAM, *Professor*

Coffee 16:20 - 16:50 Venue: Avery Room at the Student Center

Special Invited Session 3 Jae-Kwang Kim, Jean Opsomer

Venue: SC Ball A

Chair : Sanjay CHAUDHURI, University of Nebraska-Lincoln

16:50 **Multiple bias calibration for adjusting selection bias of voluntary samples using data integration** [Abstract 132]

Jae-kwang KIM, *Iowa State University*
Zhonglei WANG, *Xiamen University*
Shu YANG, *North Carolina State University*

17:30 **Fitting Classification Trees to Complex Survey Data** [Abstract 175]

Jean OPSOMER, *Westat*
Minsun RIDDLES, *Westat*

01.E1.I21 Bayesian Modeling and Computation

Venue: SC Ball B

Chair and Organizer : Joyee GHOSH, The University of Iowa

16:50 **Bayesian finite mixture of regression analysis for cancer based on histopathological imaging–environment interactions** [Abstract 119]

Yunju IM, *University of Nebraska Medical Center*

17:15 **Individualized Inference in Bayesian Quantile Directed Acyclic Graphical Models** [Abstract 28]

Anindya BHADRA, *Purdue University*
Ksheera SAGAR, *Purdue University*
Yang NI, *Texas A&M University*
Veera BALADANDAYUTHAPANI, *University of Michigan*

17:40 **Online Bayesian Variable Selection for Streaming Data** [Abstract 91]

Joyee GHOSH, *The University of Iowa*
Aixin TAN, *The University of Iowa*

01.E1.I22 The Society for Clinical Trials: 45 Years of Collaboration between Statisticians, Clinicians, Ethicists, and Trialists

Venue: SC Ball C

Chair : Jiabu YE, Merck Research Labs

Organizer : Rick CHAPPELL, University of Wisconsin **Sponsor:** The Society for Clinical Trials

16:50 **What is the SCT and what's in it for you - a historical journey** [Abstract 53]

Rick CHAPPELL, *University of Wisconsin*

17:15 **How to learn more about practical aspects of clinical trials** [Abstract 81]

Dixie ECKLUND, *University of Iowa*

17:40 **Hot off the presses - The SCT Trial of the Year** [Abstract 129]

Amarjot KAUR, *Merck Research Labs*

01.E1.I23 Statistical Challenges and Opportunities in data-driven drug development

Venue: MZ 122

Chair and Organizer : Bharani DHARAN, Novartis Pharmaceuticals

16:50 **Optimizing Treatment Allocation in Randomized Clinical Trials by Leveraging Baseline Covariates** [Abstract 245]

Zhiwei ZHANG, *Gilead Sciences*
Wei ZHANG, *Chinese Academy of Sciences*
Aiyi LIU, *National Institutes of Health*

17:15 **Using External Control Arm to Benchmark Time-to-Event Outcomes in Single-Arm Trials: A Case Study on Triple Negative Breast Cancer Patients** [Abstract 77]

Abhishek DUBEY, *Bristol Myers Squibb*
Arun KUMAR, *Bristol Myers Squibb*
Kaushal MISHRA, *Bristol Myers Squibb*
Kalyanee VIRASWAMI-APPANNA, *Bristol Myers Squibb*
Armand CHOUZY, *Bristol Myers Squibb*
Ram TIWARI, *Bristol Myers Squibb*

17:40 **Evaluation of methods for analyzing the impact of crossover in Oncology trials** [**Abstract 198**]

Abhijoy SAHA, *Eli Lilly and Company*

01.E1.I24 Statistical Inference on High Dimensional Data

Venue: MZ 226

Chair : Nairanjana DASGUPTA, Washington State University
Organizer : Swarnita CHAKRABORTY, Washington State University

16:50 **Inference for Change Points in High Dimensional Mean Shift Models** [**Abstract 127**]

Abhishek KAUL, *Washington State University*

17:15 **On the Realization Problem of Tail Dependence Matrices** [**Abstract 215**]

Narikadu D. SHYAMALKUMAR, *University of Iowa*

17:40 **The power of r-power** [**Abstract 69**]

Nairanjana DASGUPTA, *Washington State University*

01.E1.I25 Empirical Bayes Methodology

Venue: MZ 235

Chair and Organizer : Adityanand GUNTUBOYINA, University of California Berkeley

16:50 **Empirical partially Bayes multiple testing and compound Chi-square decisions** [**Abstract 206**]

Bodhisattva SEN, *Columbia University*
Nikolaos IGNATIADIS, *Columbia University*

17:15 **The edge of discovery: Controlling the local false discovery rate at the margin** [**Abstract 219**]

Jake SOLOFF, *University of Chicago*
Daniel XIANG, *University of Chicago*
William FITHIAN, *University of California, Berkeley*

17:40 **Empirical Bayes estimation: When does g-modeling beat f-modeling in theory (and in practice)?** [**Abstract 236**]

Yihong WU, *Yale University*
Yandi SHEN, *UChicago*

Friday June 2

Registration

Venue: `Friedhoff Lobby`

Time : **8:00 - 18:00**

Plenary Lecture 1 Doug Nychka

Venue: **Friedhoff A**

Chair : Soutir BANDYOPADHYAY, Department of Applied Mathematics and Statistics, Colorado School of Mines

9:00 **Fast methods for conditional simulation, the key to spatial inference.** [Abstract 173]

Doug NYCHKA, *Colorado School of Mines*

Plenary Lecture 2 John Abowd

Venue: **Friedhoff A**

Chair : Anindya ROY, University of Maryland Baltimore County

10:00 **Differential Privacy and the Overall Privacy of 2020 Census Data** [Abstract 1]

John ABOWD, *Cornell University and U.S. Census Bureau*

Coffee 11:00 - 11:30 Venue: Green Center Lobby

Poster Session

Venue: **Friedhoff B**

Time : **11:00 - 14:00**

1. **Covariance-Based Clustering for Classification** [Abstract 9]

   Theophilus ANIM BEDIAKO, *South Dakota State University*

2. **Temporal Downscaling for Solar Radiation Using a Diurnal Template Model** [Abstract 15]

   Maggie, D BAILEY, *Colorado School of Mines*

3. **Performance guarantees of Spectral clustering using g-distance and Longest leg path distance** [Abstract 25]

   Sabyasachi BERA, *University of Minnesota*

4. **Scalable Community Detection in Massive Networks via Predictive Assignment** [Abstract 29]

   Subhankar BHADRA, *North Carolina State University*

5. **Structured Dynamic Pricing: Optimal Regret in a Global Shrinkage Model** [Abstract 33]

   Rashmi Ranjan BHUYAN, *University of Southern California*

6. **A prescreening methodology for the use of likelihood ratios with subpopulation structures in the alternative source population** [Abstract 34]

   Dylan, D BORCHERT, *South Dakota State University*

7. **A general framework for regression with mismatched data** [Abstract 42]

   Priyanjali BUKKE, *George Mason University*

8. **Subsampling Based Community Detection for Large Networks** [Abstract 47]

   Sayan CHAKRABARTY, *University of Illinois at Urbana Champaign*

9. **On the Analysis of Large Scale Observational Streaming Data** [Abstract 51]

   Aleena CHANDA, *University of Nebraska-Lincoln*

10. **Evaluation of small-area estimators and associated MSPE estimators under model misspecification** [Abstract 61]

    Yuting CHEN, *University of Maryland, College Park*

11. **Using continuous methane measurements for inventory development on oil and gas sites: three case studies** [Abstract 63]

    William, S DANIELS, *Colorado School of Mines*

12. **Multi-scale Genome-wide Mediation Analysis (M-GMAS) in Twin Imaging Studies** [Abstract 65]

    Anisha DAS, *Florida State University*

13. **Blocked Gibbs sampler for hierarchical Dirichlet processes** [Abstract 67]

    Snigdha DAS, *Department of Statistics, Texas A&M University*

14. **Hierarchical Bayes estimation of small area proportions using statistical linkage of disparate data sources** [Abstract 68]

    Soumojit DAS, *University of Maryland, College Park*

15. **Sketched Gaussian Processes: A Strategy for Predictive Inference for High-Dimensional Features** [Abstract 85]

    Samuel F. GAILLIOT, *Texas A&M University*

16. **Likelihood-based spatiotemporal forecasting of burned area due to wildfire** [Abstract 86]

    Indrila GANGULY, *North Carolina State University*

17. **The envelope distribution of a complex Gaussian random variable** [Abstract 89]

    SATTWIK GHOSAL, *IOWA STATE UNIVERSITY*

18. **Employing Tensor Regression for Analyzing the Effect of Alcohol on the Brain** [Abstract 95]

    Clarissa L. GIEFER, *South Dakota State University*

19. **Application of Gaussian Mixture Models to Simulated Additive Manufacturing** [Abstract 110]

    Jason HASSE, *South Dakota State University*
    Semhar MICHAEL, *South Dakota State University*
    Anamika PRASAD, *Florida International University*

20. **Methane emission detection, localization, and quantification using continuous point-sensors on oil and gas facilities** [Abstract 121]

    Meng JIA, *Colorado School of Mines*

21. **Finite mixture of regression models for censored data based on the skew-t distribution** [Abstract 178]

    Jiwon PARK, *University of Connecticut*

22. **A Matrix Ensemble Kalman Filter-based Multi-arm Neural Network to Adequately Approximate Deep Neural Networks** [Abstract 184]

    Ved PIYUSH, *Department of Statistics, University of Nebraska - Lincoln*

23. **Estimation of finite population proportions for small areas – a statistical data integration approach** [Abstract 205]

    Aditi SEN, *University of Maryland*

24. **Predicting Rebel Movements in Mexico: A Machine Learning Approach** [Abstract 211]

    Harsh Hemant SHAH, *University of Minnesota*

25. **Predicting the anti-government conflicts based on Spatio-Temporal Geo-Political data from Mexico** [Abstract 222]

    Vishal SUBEDI, *University of Minnesota*

Lunch 12:30 - 14:00 Venue: Green Center Lobby

Bahadur Memorial Lecture Amarjit Budhiraja

Venue: Friedhoff A

Chair : Subhashish GHOSHAL, North Carolina State University

14:00 **Large Deviations and Stochastic Control** [Abstract 41]

Amarjit BUDHIRAJA, *University of North Carolina Chapel Hill*

Plenary Lecture 3 Aarti Singh

Venue: Friedhoff A

Chair : Ansu CHATTERJEE, University of Minnesota

15:00 **Misspecification and Calibration effects in Sequential decision making** [Abstract 216]

Aarti SINGH, *Carnegie Mellon University*

Coffee 16:00 - 16:30 Venue: Green Center Lobby

Plenary Lecture 4 Paul Albert

Venue: Friedhoff A

Chair : Anindya ROY, University of Maryland Baltimore County

16:30 **Innovative Applications of Hidden Markov models in Cancer Data Science** [Abstract 5]

Paul ALBERT, *National Cancer Institute*

Banquet and Awards Ceremony 18:30 - 21:30 Venue: Friedhoff A

Saturday June 3

Registration

Venue: MZ Atrium

Time : **8:00 - 18:00**

Special Invited Session 4 Frank Bretz, Satrajit Roychoudhury

Venue: MZ 126

Chair : Bharani DHARAN, Novartis Pharmaceuticals

9:00 **Bringing statistical innovation into pharmaceutical drug development: Closed MCP-Mod for pairwise comparisons of several doses with a control** [Abstract 38]

Frank BRETZ, *Novartis*

9:40 **Dynamic enrichment of Bayesian small sample, sequential, multiple assignment randomized trial (snSMART) design using natural history data: A case study from Duchenne muscular dystrophy** [Abstract 193]

Satrajit ROYCHOUDHURY, *Pfizer Inc.*

Special Session 2 Toward Enhanced Collaboration among Statisticians of Indian and African Descent in North America

Venue: MZ 226

Chair : Alemayehu DEMISSIE, VP & Head of the Statistical Research and Data Science Center at Pfizer Inc.

Organizer : Yehenew KIFLE, Department of Mathematics and Statistics, University of Maryland Baltimore County (UMBC)

9:00 **Opening Remarks**

9:05 **Ridge–Type Shrinkage Estimators in Low and High Dimensional Beta Regression Model with Applications** [Abstract 117]

Abdulkadir HUSSEIN, *University of Windsor*
Abdulkadir HUSSEIN, *University of Windsor, Canada*
Reza BELAGHI, *University of Windsor, Canada*
Yasin ASAR, *Necmettin Erbakan University, Konya, Turkey*
AUTHOR 4: S. E. AHMED, BROCK UNIVERSITY, CANADA,

9:30 **Modeling heterogeneity in hierarchically structured data for source identification problems** [Abstract 161]

Semhar MICHAEL, *South Dakota State University*
Andrew SIMPSON, *South Dakota State University*
Dylan BORCHERT, *South Dakota State University*
Christopher SAUNDERS, *South Dakota State University*
COAUTHOR 4 - LARRY TANG, LIANSHENG.TANG@UCF.EDU, UNIVERSITY OF CENTRAL FLORIDA,

9:55 **Comparison of local powers of some exact tests for a common normal mean with unequal variances** [Abstract 131]

Yehenew KIFLE, *Department of Mathematics and Statistics, University of Maryland Baltimore County (UMBC)*
Bimal SINHA, *Department of Mathematics and Statistics, University of Maryland, Baltimore County (UMBC)*
Alain MOLUH, *Department of Mathematics and Statistics, University of Maryland, Baltimore County (UMBC)*

03.M1.I26 Innovative Methods for Complex Clinical Trials

Venue: MZ 122

Chair and Organizer : Amarjot KAUR, Merck Research Labs

9:00 **The Scale Transformed Power Prior for Time-To-Event Data** [Abstract 118]

Joseph IBRAHIM, *University of North Carolina*
Ethan ALT, *University of North Carolina*
Xinxin CHEN, *University of North Carolina*
Matthew PSIODA, *Glaxo-Smith-Kline (GSK)*
BRADY NIFONG OF GSK IS AN ADDITIONAL CO-AUTHOR,

9:25 **Exploratory subgroup identification in the heterogeneous Cox model: A relatively simple procedure** [Abstract 143]

Leon LARRY, *Merck*

9:50 **LEAP: The latent exchangeability prior for borrowing information from historical data** [Abstract 6]

Ethan ALT, *University of North Carolina at Chapel Hill*
Xiuya CHANG, *University of North Carolina at Chapel Hill*
Xun JIANG, *Amgen, Inc.*
Qing LIU, *Amgen, Inc.*

03.M1.I27 Recent advances in high-dimensional functional data analysis

Venue: MZ 222

Chair : Priyam DAS, Virginia Commonwealth University
Organizer : Tanujit DEY, Harvard Medical School

9:00 **Bayesian Shrinkage Kernel Regression for joint selection of microbiome data** [Abstract 242]

Liangliang ZHANG, *Case Western Reserve University*

9:25 **Determining PET brain activity using a Bayesian spatial model In Alzheimer's disease** [Abstract 243]

Lijun ZHANG, *Case Western Reserve University*

9:50 **Multivariate functional data clustering using adaptive density peak detection** [Abstract 231]

Xiaofeng WANG, *Cleveland Clinic*

03.M1.I28 Statistical Education, Data Science, and Psychometrics

Venue: MZ 235

Chair : Megan HEYMAN, Rose-Hulman Institute of Technology
Organizer : Ansu CHATTERJEE, University of Minnesota

9:00 **Frequent Use of Authentic Assessments in the Statistics Classroom** [Abstract 114]

Megan HEYMAN, *Rose-Hulman Institute of Technology*

9:25 **Statistics and Data Science at Instacart and Google** [Abstract 112]

Tim HESTERBERG, *Instacart*

9:50 **How Advanced Statistical and Data Science Methods Are Reshaping Next-generation Psychometrics** [Abstract 218]

Sandip SINHARAY, *Educational Testing Service*

03.M1.I29 Causal Inference Innovation and Application in Real World Evidence

Venue: MZ 322

Chair and Organizer : Yi HUANG, University of Maryland, Baltimore County

Sponsor: International Chinese Statistical Association

9:00 **Leveraging Real-World Data and Real-World Evidence in Clinical Trial Design and Analysis and its Causal Implications** [Abstract 226]

Chenguang WANG, *Senior Director, Regeneron Pharmaceuticals*

9:25 **Statistical Analysis of COVID-19 Impacted Bioequivalence Study Data** [Abstract 135]

Martin KLEIN, *FDA*

9:50 **Statistical Considerations for Externally Controlled Studies using Real World Data with Small Sample Size** [Abstract 212]

Gaurav SHARMA, *Takeda Pharmaceuticals*
Jo GAO, *Takeda Pharmaceuticals*

03.M1.I30 Evolving Networks: Theory and Practice

Venue: MZ 326

Chair and Organizer : Avanti ATHREYA, Johns Hopkins University

9:00 **Estimating network-mediated causal effects via spectral embeddings** [Abstract 145]

Keith LEVIN, *University of Wisconsin-Madison*
Alex HAYES, *University of Wisconsin-Madison*
Mark FREDRICKSON, *University of Michigan*

9:25 **ECoHeN: A Hypothesis Testing Framework for Extracting Communities from Heterogeneous Networks** [Abstract 83]

Bailey FOSDICK, *Colorado School of Public Health*
Connor GIBBS, *Colorado State University*
James WILSON, *University of San Francisco*

9:50 **Anomalous clique detection and identification in inhomogeneous networks** [Abstract 208]

Srijan SENGUPTA, *North Carolina State University*
Subhankar BHADRA, *North Carolina State University*

03.M1.I31 Recent Advances in Biostatistics and Bioinformatics

Venue: MZ 335

Chair : Tusharkanti GHOSH, Colorado School of Public Health
Organizer : Anindya ROY, University of Maryland Baltimore County

9:00 **A robust Kernel Machine framework for assessing differential expression of multi sampled single cell data** [Abstract 94]

Tusharkanti GHOSH, *Colorado School of Public Health*
Debashis GHOSH, *Colorado School of Public Health*

9:25 **A Curious Distribution** [Abstract 189]

Marepalli RAO, *University of Cincinnati*
Zhaochong YU, *University of Cincinnati*
Neelakshi CHATTERJEE, *University of Cincinnati*
NONE,

9:50 **Multiply-robust estimation of causal treatment effect on a binary outcome with integrated information from secondary outcomes** [**Abstract 230**]

Ming WANG, *Case Western Reserve University*
Chixiang CHEN, *University of Maryland*
Shuo CHEN, *University of Maryland*
Qi LONG, *University of Pennsylvania*
THERE IS ANOTHER CO-AUTHOR: SUDESHNA DAS FROM MASSACHUSETTS GENERAL HOSPITAL, HARVARD MEDICAL SCHOOL,

Coffee 10:20 - 10:50 Venue: MZ Atrium

Special Invited Session 5 Rahul Mazumder, George Michailidis

Venue: MZ 126

Chair : Sanjay CHAUDHURI, University of Nebraska-Lincoln

10:50 **Computational Lenses in learning combinatorial structures in statistics and machine learning.** [**Abstract 158**]

Rahul MAZUMDER, *MIT Sloan School of Management*

11:30 **Statistical models for mixed frequency data and their applications in forecasting economic indicators** [**Abstract 162**]

George MICHAILIDIS, *UCLA*

Panel Discussion 2 Leadership and Collaboration across Cultures

Venue: MZ 326

Organizers : *Mark OTTO, U.S. Fish and Wildlife Service* and *Amarjot KAUR, Merck Research Labs*

- **Eric VANCE**, University of Colorado, Boulder
- **Christine GAUSE**, Merck
- **Tim HESTERBERG**, Instacart
- **Donsig JANG**, NORC

(*Representing KISS*)

- **Ivan CHAN**, Bristol Myers Squibb

(*Representing ICSA*)

03.M2.I32 Model-based statistical learning: Method and applications

Venue: MZ 122

Chair : Jae-kwang KIM, Iowa State University
Organizer : Gyuhyeong GOH, Kansas State University

Sponsor: Korean International Statistical Association

10:50 **Bayesian Network Meta-Regression for Aggregate Ordinal Outcomes with Imprecise Categories** [**Abstract 108**]

Yeongjin GWON, *University of Nebraska Medical Center*
Ming-Hui CHEN, *University of Connecticut*
Joseph IBRAHIM, *University of North Carolina*

11:15 **Model Selection in Data Integration** [**Abstract 197**]

Takumi SAEGUSA, *University of Maryland*

11:40 Bayesian model-based synthetic control methods [Abstract 96]

Gyuhyeong GOH, *Kansas State University*
Jisang YU, *Kansas State University*

03.M2.I33 Recent Advances in Time Series and Functional Data Analysis

Venue: MZ 222

Chair : Yi HUANG, University of Maryland, Baltimore County
Organizer : Guoqing DIAO, Milken Institute School of Public Health, George Washington University

Sponsor: International Chinese Statistical Association

10:50 Frequency Band Analysis of Nonstationary Multivariate Time Series [Abstract 39]

Scott BRUCE, *Texas A&M University*
Raanju SUNDARARAJAN, *Southern Methodist University*
Scott BRUCE, *Texas A&M University*

11:15 Detection of Structural Breaks in Non-stationary Spatial Random Field [Abstract 14]

Pramita BAGCHI, *Department of Statistics, George Mason University*

11:40 Fast Generalized Functional Principal Components Analysis [Abstract 235]

Julia WROBEL, *Department of Biostatistics and Informatics, Colorado School of Public Health*

03.M2.I34 Statistical inference for complex data structures

Venue: MZ 235

Chair : Nilanjan CHAKRABORTY, Washington University in Saint Louis
Organizer : Soumendranath LAHIRI, Washington university of St. Louis

10:50 A Hybrid Empirical Likelihood for Time Series [Abstract 172]

Dan NORDMAN, *Iowa State University*
Haihan YU, *Iowa State University*
Mark KAISER, *Iowa State University*

11:15 On Estimation of Function-on-function Regression Coefficients with Brownian Berkson Errors [Abstract 185]

Paramahansa PRAMANIK, *University of South Alabama*

11:40 Change point detection and localization in a panel of densities [Abstract 137]

Piotr KOKOSZKA, *Colorado State University*

03.M2.C1 Contributed Session

Venue: MZ 226

Chair : Satya Ravi K. SIDDANI,

10:50 A Powerful Detection Rule in High Dimensional Mediation Hypothesis Testing [Abstract 192]

Asmita ROY, *Texas A&M University*
Xianyang ZHANG, *Texas A&M University*

11:05 Data Adaptive Covariate Balancing for Causal Effect Estimation for High Dimensional Data [Abstract 71]

Simion DE, *Biostatistics PhD Student, University of Minnesota*
Jared HULING, *Assistant Professor,University of Minnesota*

11:20 Estimating Prediction Error for Functional Time Series [Abstract 30]

Samayita BHATTACHARJEE, *University of California, Davis*
Alexander AUE, *University of California, Davis*
Prabir BURMAN, *University of California, Davis*

11:35 Flexible Tree-based Model for Extreme Events [Abstract 54]

Suneel Babu CHATLA, *University of Texas at El Paso, Texas*
Galit SHMUELI, *National Tsing Hua University, Hsinchu, Taiwan*

11:50 CMPLE to decode Photosynthesis using MM algorithm [Abstract 58]

Abhijnan CHATTOPADHYAY, *Postdoctoral Fellow, National Institute of Environmental Health Science, National Institute of Health*
David Mark KRAMER, *Professor, Plant Research Lab, Michigan State University*
Tapabrata MAITI, *Professor, Department of Statistics, Michigan State University*
Samiran SINHA, *Professor, Department of Statistics, Texas A & M University*

Lunch 12:10 - 13:30 Venue: MZ Atrium

Short Course 2 Workshop on Machine Learning: ML Algorithms - Explainability, Diagnostics and Model Validation

Venue: MZ 326
Time : **13:30 - 18:00**

- **Anwesha BHATTACHARYA**, Wells Fargo
- **Agus SUDJIANTO**, Wells Fargo

Special Invited Session 6 Hal Stern, Purnamrita Sarkar

Venue: MZ 126
Chair : Sandip SINHARAY, Educational Testing Service

13:30 Statistics and the Fair Administration of Justice: Assessing Bloodstain Pattern Evidence [Abstract 220]

Hal STERN, *University of California Irvine*
Tong ZOU, *University of California Irvine*

14:10 Bootstrapping the Error of Oja's Algorithm [Abstract 203]

Purnamrita SARKAR, *University of Texas at Austin*

03.A1.I35 Statistical analysis of random graphs: theory and applications

Venue: MZ 122
Chair and Organizer : Srijan SENGUPTA, North Carolina State University

13:30 Multiple Hypothesis Testing To Estimate The Number of Communities in Sparse Stochastic Block Models [Abstract 120]

Chetkar JHA, *Washington University in St. Louis*
Li MINGYAO, *Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania*
Ian BARNETT, *Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania*

13:55 On network modularity statistics in connectomics and schizophrenia [Abstract 45]

Joshua CAPE, *University of Wisconsin*
Anirban MITRA, *University of Pittsburgh*
Konasale PRASAD, *University of Pittsburgh*

14:20 **Attributed random networks: local weak limits, PageRank and sampling fairness** [**Abstract 21**]

Sayan BANERJEE, *University of North Carolina, Chapel Hill*
Nelson ANTUNES, *Center for Computational and Stochastic Mathematics,University of Lisbon*
Shankar BHAMIDI, *UNC Chapel Hill*
Vladas PIPIRAS, *UNC Chapel Hill*

03.A1.I36 Recent advances in nonparametric and object-oriented regressions in complex structured data

Venue: MZ 222

Chair and Organizer : Samuel F. GAILLIOT, Texas A&M University

13:30 **Multi-object Data Integration in the Study of Primary Progressive Aphasia** [**Abstract 107**]

Rene GUTIERREZ, *Texas A&M University*

13:55 **Bayesian modeling with derivative Gaussian processes of event-related potentials** [**Abstract 240**]

Cheng-Han YU, *Marquette University*
Meng LI, *Rice University*
Marina VANNUCCI, *Rice University*

14:20 **Distributed Inference and Data Compression: A Tale of Two Techniques** [**Abstract 103**]

Rajarshi GUHANIYOGI, *Texas A&M University*

03.A1.I37 Computational complexities of spatio-temporal modeling in Environmental Sciences

Venue: MZ 226

Chair and Organizer : Indranil SAHOO, Virginia Commonwealth University

13:30 **Multivariate Spatial Prediction of Air Pollutants** [**Abstract 97**]

Wenlong GONG, *University of Houston - Downtown*
Brian REICH, *North Carolina State University*
Howard CHANG, *Emory University*

13:55 **Overcoming computational complexities of large-scale spatial modeling with Nearest Neighbor Gaussian Processes using the BRISC R-package** [**Abstract 199**]

Arkajyoti SAHA, *University of Washington, Department of Statistics*
Abhirup DATTA, *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health*

14:20 **Modeling High-Dimensional Spatial Covariance Structures for Climate Processes using Physically-Informed Basis Functions** [**Abstract 24**]

Samuel BAUGH, *Lawrence Berkeley National Laboratory*
Samuel BAUGH, *Lawrence Berkeley National Lab*
Karen MCKINNON, *UCLA*
Mark RISSER, *Lawrence Berkeley National Lab*

03.A1.I38 Robust network inference

Venue: MZ 235

Chair and Organizer : Avanti ATHREYA, Johns Hopkins University

13:30 **Joint spectral clustering in multilayer networks** [**Abstract 10**]

Jesús ARROYO, *Texas A&M University*
Joshua AGTERBERG, *University of Pennsylvania*
Zachary LUBBERTS, *Johns Hopkins University*

13:55 Incorporating network side information into supervised learning [Abstract 207]

Subhabrata SEN, *Harvard University*
Sagnik NANDY, *University of Pennsylvania*
Subhabrata SEN, *Harvard University*

03.A1.I39 Recent methods in factor analysis

Venue: MZ 322

Chair : Sayar KARMAKAR, University of Florida
Organizer : Somak DUTTA, Iowa State University

13:30 Matrix-free maximum likelihood estimation of Gaussian factor models [Abstract 80]

Somak DUTTA, *Iowa State University*
Ranjan MAITRA, *Iowa State University*
Fan DAI, *Michigan Technological University*

13:55 Factor Analysis of Data with Incomplete Records [Abstract 62]

Fan DAI, *Michigan Technological University*
Somak DUTTA, *Iowa State University*
Ranjan MAITRA, *Iowa State University*

14:20 Exploratory Factor Analysis for Data on a Sphere [Abstract 151]

Ranjan MAITRA, *Iowa State University*
Fan DAI, *Michigan Technological University*
Karin DORMAN, *Iowa State University*
Somak DUTTA, *Iowa State University*

03.A1.I40 Recent Developments in High-dimensional Bayesian Modeling and Inference

Venue: MZ 335

Chair and Organizer : Nilabja GUHA, University of Massachusetts Lowell

13:30 Change-point detection in high-dimension [Abstract 100]

Nilabja GUHA, *University of Massachusetts Lowell*
Jyotishka DATTA,

13:55 On the Convergence of Coordinate Ascent Variational Inference [Abstract 31]

Anirban BHATTACHARYA, *Texas A&M University*
Debdeep PATI, *Texas A&M University*
Yun YANG, *University of Illinois at Urbana-Champaign*

14:20 Title: Adaptive finite element type decomposition of Gaussian random fields [Abstract 179]

Debdeep PATI, *Texas A&M University*
Jaehoan KIM, *Texas A&M University*
Anirban BHATTACHARYA, *Texas A&M University*

Short Break 14:50 - 15:00

Memorial Session 2 In Memorium: Professor Krishna Athreya

Venue: MZ 126

Chair : Anindya ROY, University of Maryland Baltimore County

15:00 **Remarks**

15:10 **Ancestral Inference for Bellman Harris process.** [Abstract 225]

Anand VIDYASHANKAR, *George Mason University*

15:45 **Quadratic Prediction Methodology and Calibration of Prediction Intervals Based on Subsampling** [Abstract 142]

Soumendranath LAHIRI, *Washington university of St. Louis*

03.A2.I41 Recent advances in anytime-valid sequential inference

Venue: MZ 122

Chair and Organizer : Aaditya RAMDAS, Carnegie Mellon University

15:00 **Multi Armed Bandits and Empirical Likelihood Confidence Intervals** [Abstract 3]

Shubhada AGRAWAL, *Georgia Institute of Technology*
Sandeep JUNEJA, *Tata Institute of Fundamental Research, Mumbai*
Wouter KOOLEN, *CWI, Amsterdam*

15:25 **Anytime-Valid Confidence Sequences in an Enterprise A/B Testing Platform** [Abstract 217]

Ritwik SINHA, *Adobe Research*

15:50 **Sequential change detection via backward confidence sequences** [Abstract 214]

Shubhanshu SHEKHAR, *Carnegie Mellon University*
Aaditya RAMDAS, *Department of Statistics and Data Science, Carnegie Mellon University*

03.A2.I42 New Developments in Survival and Longitudinal Analysis

Venue: MZ 222

Chair and Organizer : Sandip SINHARAY, Educational Testing Service

15:00 **Combining efficacy, safety, and patient reported outcomes in cancer drug development – a late phase example** [Abstract 2]

Suddhasatta ACHARYYA, *Gilead Sciences*

15:25 **Assessing Contribution of Treatment Phases through Tipping Point Analyses via Counterfactual Elicitation Using Rank Preserving Structural Failure Time Models** [Abstract 75]

Jyotirmoy DEY, *Regeneron Pharmaceuticals, Inc.*
Sudipta BHATTACHARYA, *Takeda Pharmaceuticals, Inc.*

15:50 **Piecewise Random-Effects Models for Segmented Longitudinal Trends** [Abstract 136]

Nidhi KOHLI, *University of Minnesota*

03.A2.I43 AI/ML Applications in the Industry

Venue: MZ 226

Chair : Abhijoy SAHA, Eli Lilly and Company
Organizer : Sai Kumar POPURI, Bed Bath & Beyond Inc.

15:00 **TSEC: a framework for online experimentation under experimental constraints** [Abstract 154]

Simon MAK, *Duke University*

15:25 **AI/ML in retail digital transformation** [Abstract 224]

Shirish TATIKONDA, *Walmart Data Ventures*

15:50 **Observing Relations in Mixed Corpora through Abstractive Summarization with Dimension Reduction: A Topic Modeling Approach** [Abstract 164]

R. Cole MOLLOY, *JHU/APL*

03.A2.I44 Recent developments in small area statistics and applications

Venue: MZ 235

Chair : Gauri DATTA, Univ of Georgia and US Census Bureau
Organizer : IISA,

15:00 **On an Empirical likelihood-based Estimator for Complex Survey Data** [Abstract 59]

Sanjay CHAUDHURI, *University of Nebraska-Lincoln*

15:25 **RESPONSE MODEL SELECTION IN CASE OF NOT MISSING AT RANDOM NONRESPONSE WITH APPLICATION TO REAL DATA** [Abstract 223]

Michael SVERCHKOV, *Bureau of Labor Statistics*

15:50 **A Pseudo-Bayesian Approach to Small Area Estimation Using Spatial Models** [Abstract 70]

Gauri DATTA, *Univ of Georgia and US Census Bureau*
Jiacheng LI, *Univ of Georgia*

03.A2.I45 Theory of Cross Validation

Venue: MZ 335

Chair : Adityanand GUNTUBOYINA, University of California Berkeley
Organizer : Sabyasachi CHATTERJEE, University of Illinois at Urbana Champaign

15:00 **Asymptotics of cross validation** [Abstract 13]

Morgane AUSTERN, *Harvard*
Wenda ZHOU, *NYU/Flatiron*

15:25 **Understanding and approximating leave-one-out cross validation under high-dimensional asymptotics** [Abstract 155]

Arian MALEKI, *Columbia University*
Arnab AUDDY, *Columbia University*
Haolin ZOU, *Columbia University*
Kamiar RAHNAMA RAD, *City University of New York*

15:50 **A Theoretically Tractable Cross Validation Framework for Signal Denoising** [Abstract 56]

Sabyasachi CHATTERJEE, *University of Illinois at Urbana Champaign*

Coffee 16:20 - 16:50 Venue: MZ Atrium

Memorial Session 3 In Memorium: Professor Krishna Athreya

Venue: MZ 126

Chair : Saonli BASU, University of Minnesota

16:50 Connections between stick-breaking measures and Markov chains [Abstract 209]

Sunder SETHURAMAN, *University of Arizona*

17:25 Limiting spectral distribution of random matrices with independent entries [Abstract 35]

Arup BOSE, *Indian Statistical Institute*
Priyanka SEN, *Indian Statistical Institute*
Arusharka SEN, *Concordia University*
Koushik SAHA, *koushiksaha877@gmail.com*

03.E1.I46 Topics in high-dimensional statistics

Venue: MZ 122

Chair : Krishna BALASUBRAMANIAN, UC Davis
Organizer : Po-Ling LOH, University of Cambridge

16:50 High-dimensional Central Limit Theorems for Linear Functionals of Online Least-Squares SGD [Abstract 16]

Krishna BALASUBRAMANIAN, *UC Davis*

17:15 Spectral Universality in Regularized Linear Regression with Nearly Deterministic Designs [Abstract 78]

Rishabh DUDEJA, *Harvard University*
Subhabrata SEN, *Harvard University*
Yue M. LU, *Harvard University*

17:40 Debiasing in missing data models with inaccurate nuisance parameters [Abstract 46]

Michael CELENTANO, *UC Berkeley*

03.E1.I47 Modern semiparametric methods for biomedical data

Venue: MZ 222

Chair and Organizer : Moumita KARMAKAR, Texas A&M University

16:50 Nonparametric estimation of the age-at-onset distribution from a cross-sectional sample [Abstract 157]

Soutrik MANDAL, *NYU Grossman School of Medicine*
Jing QIN,
Ruth PFEIFFER,

17:15 A Bayesian semiparametric model for variable selection in compositional data [Abstract 200]

Satabdi SAHA, *The University of Texas MD Anderson Cancer Center*
Christine PETERSON, *The University of Texas MD Anderson Cancer Center*

17:40 Country-specific Estimation of Verbal Autopsy Misclassification in Improving Global Mortality Surveillance [Abstract 186]

Sandipan PRAMANIK, *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health*
Abhirup DATTA, *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health*
Scott ZEGER, *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health*

03.E1.I48 Shape-constrained Inference

Venue: MZ 226

Chair : Bodhisattva SEN, Columbia University
Organizer : Adityanand GUNTUBOYINA, University of California Berkeley

16:50 Covariance estimation with nonnegative partial correlations [Abstract 106]

Adityanand GUNTUBOYINA, *University of California Berkeley*
Jake A. SOLOFF, *University of Chicago*
Michael I. JORDAN, *University of California Berkeley*

17:15 Inference on contrasts of monotone functions [Abstract 233]

Ted WESTLING, *Department of Mathematics and Statistics, University of Massachusetts Amherst*
Yujian WU, *University of Massachusetts Amherst*
Eric MORENZ, *University of Washington*
Marco CARONE, *University of Washington*

17:40 On the Variance and Admissibility of Empirical Risk Minimization [Abstract 140]

Gil KUR, *MIT*
Alexander RAKHLIN, *MIT*
Eil PUTTERMAN, *Tel Aviv University*

03.E1.I49 Statistical modeling and inference for wildland fires

Venue: MZ 235

Chair and Organizer : Srijan SENGUPTA, North Carolina State University

16:50 Deep Hierarchical Generalized Transformation Models with Application to Wildfires [Abstract 36]

Jonathan BRADLEY, *Florida State University*

17:15 Fire-Atmosphere Coupling Implications for Wildland Fire Modeling and Decision Making [Abstract 232]

Joseph WERNE, *NorthWest Research Associates*
Joseph WERNE, *NorthWest Research Associates*
Wayne SPENCER, *Conservation Biology Institute*

03.E1.I50 Inference Problems based on Network Data

Venue: MZ 335

Chair and Organizer : Shirshendu CHATTERJEE, City University of New York

16:50 Variational Inference: Posterior Threshold Improves Network Clustering Accuracy in Sparse Regimes [Abstract 144]

Can LE, *University of California, Davis*
Xuezhen LI, *University of California, Davis*
Can LE, *University of California, Davis*

17:15 Estimation of the number of communities for sparse networks [Abstract 32]

Sharmodeep BHATTACHARYYA, *Oregon State University*
Neil HWANG, *Bronx Community College, City University of New York*
Jiarui (Sam) XU, *Meta Inc.*
Shirshendu CHATTERJEE, *City College, City University of New York*

17:40 Change point detection, estimation, and localization for network data [Abstract 57]

Shirshendu CHATTERJEE, *City University of New York*
Soumendu Sundar MUKHERJEE, *Indian Statistical Institute*
Sharmodeep BHATTACHARYYA, *Oregon State University*

Meeting 1 General Body Meeting (members only)

Venue: MZ 326

Time : **18:30 - 20:00**

- IISA general body meeting.

Sunday June 4

Registration

Venue: MZ Atrium

Time : **8:00 - 12:00**

Short Course 3 Data Analysis after Record Linkage: Sources of Error, Consequences, and Possible Solutions

Venue: MZ 326

Time : **9:30 - 12:30**

- **Martin SLAWSKI**, George Mason University
- **Emanuel BEN-DAVID**, U.S. Census Bureau
- **Priyanjali BUKKE**, George Mason University

04.M1.I51 Modern advances in Bayesian techniques

Venue: MZ 122

Chair : Sanjib BASU, UIC

Organizer : Ansu CHATTERJEE, University of Minnesota

9:00 **Bayesian Group Sparsity and Smoothing on Graphs** [Abstract 201]

Huiyan SANG, *Texas A&M University*

9:25 **Joint modeling with high dimensional longitudinal processes** [Abstract 23]

Sanjib BASU, *UIC*

9:50 **Generalized Variable Selection Algorithms for Gaussian Process Models by LASSO-like Penalty** [Abstract 74]

Dipak DEY, *University of Connecticut*
Zhiyong HU, *Microsoft New England Research and Development Center*

04.M1.I52 Mixed-effect prediction and computer experiments

Venue: MZ 222

Chair : Partha LAHIRI, University of Maryland College Park

Organizer : Ansu CHATTERJEE, University of Minnesota

9:00 **A Random-effects Approach to Regression Involving Many Categorical Predictors and Their Interactions** [Abstract 122]

Jiming JIANG, *University of California, Davis*
Hanmei SUN, *Shandong Normal University*
Jiangshan ZHANG, *University of California, Davis*

9:25 **On Estimation of the Logarithm of the Mean Squared Prediction Error of A Mixed-effect Predictor** [Abstract 171]

Thuan NGUYEN, *Oregon Health and Science University*
Jianling WANG, *Shandong University*
Yihui LUAN, *Shandong University*
Jiming JIANG, *University of California, Davis*

9:50 **An Agnostic Fay-Herriot Model For Small Area Statistics** [Abstract 55]

Ansu CHATTERJEE, *University of Minnesota*

04.M1.I53 Machine learning and statistics

Venue: MZ 226

Chair : Ramya KORLAKAI VINAYAK, UW-Madison
Organizer : Po-Ling LOH, University of Cambridge

9:00 **Estimating the fraction of anomaly points** [Abstract 166]

Debashis MONDAL, *Washington University in St Louis*

9:25 **Simple Binary Hypothesis Testing: Locally-Private and Communication-Efficient** [Abstract 182]

Ankit PENSIA, *IBM Research*
Amir ASADI, *University of Cambridge*
Varun JOG, *University of Cambridge*
Po-Ling LOH, *University of Cambridge*

9:50 **Learning from Diverse Data in Metric and Preference Learning** [Abstract 138]

Ramya KORLAKAI VINAYAK, *UW-Madison*
Gokcan TATLI,
Greg CANAL,
Blake MASON,
Rob NOWAK,

04.M1.I54 Recent developments in statistical learning theory and decision making

Venue: MZ 235

Chair and Organizer : Sakshi ARYA, Pennsylvania State University

9:00 **Cross Validation Importance Learning (CVIL)** [Abstract 238]

Chenglong YE, *University of Kentucky*
Yuhong YANG, *University of Minnesota*

9:25 **Survival Bandits** [Abstract 177]

Yinghao PAN, *University of North Carolina at Charlotte*
Eric LABER, *Duke University*
Yingqi ZHAO, *Fred Hutchinson Cancer Research Center*

9:50 **Kernel Epsilon-greedy strategy for nonparametric bandits** [Abstract 11]

Sakshi ARYA, *Pennsylvania State University*
Bharath K. SRIPERUMBUDUR, *Pennsylvania State University*

04.M1.I55 Scalable Bayesian Methods in Biology and Public Health

Venue: MZ 322

Chair : Paromita BANERJEE, JOHN CARROLL UNIVERSITY
Organizer : Tanujit DEY, Harvard Medical School

9:00 **Bayesian inference on Covid-19 transmission dynamics in India using a modified SEIR model** [Abstract 20]

Paromita BANERJEE, *JOHN CARROLL UNIVERSITY*
Kai YIN, *Case Western Reserve University*
Anirban MONDAL, *Case Western Reserve University*

9:25 **Clustering sequence data with mixture Markov chains with covariates using multiple simplex constrained optimization routine** [Abstract 66]

Priyam DAS, *Virginia Commonwealth University*
Deborshee SEN , *Amazon, India*
Debsurya DE, *Johns Hopkins University*
Tianxi CAI, *Harvard Medical School*

9:50 **Scalable Bayesian Variable Selection and Grouping for Binary and Multiclass Outcome Data** [Abstract 50]

Sounak CHAKRABORTY, *University of Missouri*

04.M1.I56 Applications of spatial methodology

Venue: MZ 335

Chair and Organizer : Rajarshi GUHANIYOGI, Texas A&M University

9:00 **Spatial scale-aware tail dependence modeling for high-dimensional spatial extremes** [Abstract 210]

Ben SHABY, *Colorado State University*

9:25 **Geostatistical capture-recapture models** [Abstract 115]

Mevin HOOTEN, *The University of Texas at Austin*

Coffee 10:20 - 10:50 Venue: MZ Atrium

04.M2.I57 Variable selection in complex data

Venue: MZ 122

Chair and Organizer : Souparno GHOSH, University of Nebraska-Lincoln

10:50 **Double-robust Bayesian variable selection and model prediction with spherically symmetric errors** [Abstract 229]

Min WANG, *University of Texas-San Antonio*

11:15 **Variable selection in quantile regression with ordinal responses** [Abstract 64]

Mai DAO, *Wichita State University*
Md Sakhawat HOSSAIN, *Texas Tech University*

11:40 **Variable selection in non-linear metric learning** [Abstract 93]

Souparno GHOSH, *University of Nebraska-Lincoln*

04.M2.I58 Inference and applications with complex surveys and small areas

Venue: MZ 222

Chair and Organizer : Ansu CHATTERJEE, University of Minnesota

10:50 **Nonprobability follow-up sample analysis: an application to SARS-Cov-2 infection prevalence estimation** [Abstract 146]

YAN LI, *UNIVERSITY OF MARYLAND AT COLLEGE PARK*
LAURA YEE, *NATIONAL INSTITUTE OF HEALTH*
SALLY HUNSBERGER, *NATIONAL INSTITUTE OF HEALTH*
BARRY GRAUBARD, *NATIONAL INSTITUTE OF HEALTH*

11:15 **A Multivariate Bayesian Hierarchical Model for Small Area Estimation of Criminal Victimization Rates in Domains Defined by Age and Gender** [Abstract 27]

Emily BERG, *Iowa State University*

04.M2.I59 Recent Advances in Statistical Inference

Venue: MZ 235

Chair : Neeraj MISRA, Indian Institute of Technology, Kanpur, India
Organizers : *Monika BHATTACHARJEE, Indian Institute of Technology, Bombay* and *Kaushik JANA, Ahmedabad University, Gujarat*

10:50 **Establishing minimax lower bound in high order tensor models for Neuroimaging** [**Abstract 19**]

Chitrak BANERJEE, *Wells Fargo Bank, NA*
Lyudmila SAKHANENKO, *Michigan State University*
David C. ZHU, *Michigan State University*

11:15 **Statistical modelling based on distributional representation of wearable data** [**Abstract 102**]

Pratim GUHA NIYOGI, *Johns Hopkins Bloomberg School of Public Health*
Vadim ZIPUNNIKOV, *Johns Hopkins Bloomberg School of Public Health*

11:40 **Estimating the Selected Treatment Effect Using a Two-Stage Adaptive Design** [**Abstract 163**]

Neeraj MISRA, *Indian Institute of Technology, Kanpur, India*
Masihuddin MASIHUDDIN, *Indian Institute of Technology, Kanpur, India*

04.M2.I60 Bayesian semiparametric and spatial models in ecology, epidemiology, and finance

Venue: MZ 322
Chair : Paromita BANERJEE, JOHN CARROLL UNIVERSITY
Organizer : Anirban MONDAL, Case Western Reserve University

10:50 **Bayesian estimation of local volatility from option pricing data** [**Abstract 165**]

Anirban MONDAL, *Case Western Reserve University*
Kai YIN, *Case Western Reserve University*

11:15 **Bayesian Semiparametric Covariate Informed Multivariate Density Deconvolution** [**Abstract 202**]

Abhra SARKAR, *The University of Texas at Austin*

11:40 **A Hierarchical Bayesian Entity Resolution Model to Improve Tree Demography Using Overlapping Lidar Scans** [**Abstract 124**]

Andee KAPLAN, *Colorado State University*
Lane DREW, *Colorado State University*
Ian BRECKHEIMER, *Rocky Mountain Biological Laboratory*

04.M2.I61 Sharp theoretical guarantees on modern machine-learning methods

Venue: MZ 335
Chair and Organizer : Nilanjana LAHA, Texas A&M

10:50 **Deep Neural Networks for Nonparametric Interaction Models with Diverging Dimension** [**Abstract 168**]

Debarghya MUKHERJEE, *Princeton University*
Sohom BHATTACHARYA, *Princeton University*
Jianqing FAN, *Princeton University*

11:15 **Policy evaluation in reinforcement learning: The impact of temporal dependence and multi-step lookahead** [**Abstract 76**]

Yaqi DUAN, *MIT*
Martin WAINWRIGHT, *MIT*

11:40 Limit theorems for high dimensional least-square online SGD [**Abstract** 88]

Promit GHOSAL, *Massachusetts Institute of Technology*
Bhavya AGRAWALLA, *Massachusetts Institute of Technology*
Krishnakumar BALASUBRAMANIAN, *University of California, Davis*
Ye HE, *University of California, Davis*

Abstracts

1. Differential Privacy and the Overall Privacy of 2020 Census Data

[Plenary Lecture 2, (page 15)]

John ABOWD, *Cornell University and U.S. Census Bureau*

Advances in computing power and optimization algorithms have made traditional statistical disclosure limitation methods, especially those used for unweighted tabulations of household and business data, increasingly vulnerable to successful confidentiality-breaching attacks. Effective mitigation of this disclosure risk, particularly for a highly granular statistical products like the Decennial Census of Population and Housing, which routinely publishes far more statistics than it has data points, requires careful evaluation and application of quantifiable and tunable confidentiality protections. This talk provides a technical overview of the differential privacy framework for the 2020 Census Disclosure Avoidance System. I examine some of the implementation constraints, policy decisons, criticisms, and lessons learned from its development and implementation.

2. Combining efficacy, safety, and patient reported outcomes in cancer drug development – a late phase example

[03.A2.I42, (page 26)]

Suddhasatta ACHARYYA, *Gilead Sciences*

The critical question, in the context of confirmatory oncology drug trials is whether we have a clear answer on the risk-benefit ratio. A comprehensive metric that combines both efficacy and safety endpoints, as well as, patient-reported quality of life endpoints could be very useful in this context. A popular approach is to use a modified overall survival type analysis, where, instead of the usual survival time, a 'quality-adjusted' version of it is analyzed. One such measure is the Q-TWiST, which may be described as the 'quality-adjusted time without symptoms and toxicity'. Originally, QTWiST was developed for the adjuvant setting but has since been extended in several directions. Here, we adapt the QTWiST methodology to our specific situation, which is a two-arm randomized phase III oncology study, in multiple myeloma. The application of this methodology to specific clinical trials often brings up analytical issues, some of which will be discussed, as they apply to our study. Keywords: survival, quality-of-life, oncology, risk-benefit

3. Multi Armed Bandits and Empirical Likelihood Confidence Intervals

[03.A2.I41, (page 26)]

Shubhada AGRAWAL, *Georgia Institute of Technology*
Sandeep JUNEJA, *Tata Institute of Fundamental Research, Mumbai*
Wouter KOOLEN, *CWI, Amsterdam*

Multi-armed bandit (MAB) is a popular framework for sequential decision-making in an uncertain environment. In the classical setup of MAB, the algorithm has access to a fixed and finite set of K unknown, independent probability distributions or arms. At each time step, having observed the outcomes of all the previous actions, the algorithm chooses one of the K arms and receives an independent sample drawn from the underlying distribution, which may be considered a reward. The algorithm's goal is either to maximize the accumulated rewards or to identify the arm with the maximum mean in as few samples as possible.

These problems are well-studied in literature, and tight lower bounds and optimal algorithms exist when the arm distributions are known to belong to simple classes of distributions such as single-parameter exponential family or distributions that have bounded support, etc. However, in practice, the distributions may not satisfy these assumptions and may even be heavy-tailed.

In this talk, we will look at techniques and algorithms for optimally solving these problems with minimal assumptions on the arm distributions. A key component of designing an optimal algorithm for MAB is constructing tight, anytime valid confidence intervals for mean. We will look at new concentration inequalities for heavy-tailed distributions, which may be of independent interest.

4. Semiparametric Regression Models for Causal Mediation Analysis with Longitudinal Data

[01.M1.I5, (page 4)]

Jeffrey ALBERT, *Case Western Reserve University*
Tanujit DEY, *Harvard Medical School and Brigham and Women's Hospital*
Hongxu ZHU, *Case Western Reserve University*
Jiayang SUN, *Ceorge Mason University*

Mediation analysis is often of interest where both the mediator and the final outcome are repeatedly measured. However, limited work has been done

for this problem and available methods are primarily based on parametric models that tend to be sensitive to model specifications. We propose semiparametric continuous time models to provide a flexible and robust approach to causal mediation analysis for longitudinal data, while also allowing these data to be unbalanced or irregular. The method uses spline linear mixed-effects models with a two-step approach to model fitting in which a predicted mediator is used as a covariate in the final outcome model. Estimated natural direct and indirect effects as a function of time are provided using an extended mediation formula and sequential ignorability assumption. We consider alternative approaches for predicting the mediator and compare properties of resulting natural direct and indirect estimators via simulations. The methodology is illustrated using harmonized data from two cohort studies to examine attention as a mediator of the effect of prenatal tobacco exposure on externalizing behavior in children.

5 . Innovative Applications of Hidden Markov models in Cancer Data Science

[Plenary Lecture 4, (page 17)]

Paul ALBERT, *National Cancer Institute*

During the past 30 years, Hidden Markov modeling (HMM) has had a big impact in the analysis of biomedical data, with a few important application areas in genomics, natural history modeling, environmental monitoring, and the analysis of longitudinal data. In cancer genomics, for example, the use of HMM has played an important role in uncovering both susceptibility (germline) and tumor progression (somatic) of cancer. In this talk, I will present a series of novel applications of HMMs in cancer epidemiology and genetics. I will describe the use of HMM to identify multiple subclones in next-generation sequences of tumor samples (Choo-Wosoba et al., Biostatistics 2021). I will also discuss the application of HMMs for characterizing the natural history of natural history of human papillomavirus and cervical precancer (Aron et al., Statistics in Medicine, 2021). Last, I will discuss the application of HMM for cancer surveillance. All three examples required interesting adaptations of standard HMM estimation that will be highlighted. Time permitting, I will talk about research opportunities using HMM in biomedical data science.

6 . LEAP: The latent exchangeability prior for borrowing information from historical data

[03.M1.I26, (page 19)]

Ethan ALT, *University of North Carolina at Chapel Hill*
Xiuya CHANG, *University of North Carolina at Chapel Hill*
Xun JIANG, *Amgen, Inc.*
Qing LIU, *Amgen, Inc.*

Author list: Ethan M. Alt, Xiuya Chang, Xun Jiang, Qing Liu, May Mo, H. Amy Xia, Joseph G. Ibrahim,

Abstract:

It is becoming increasingly popular to elicit informative priors on the basis of historical data. However, when conducting a new study, the inclusion/exclusion criteria could be different, and some of the participants in the historical data may be less relevant. Popular existing priors, including the power prior, commensurate prior, and robust meta-analytic prior provide blanket discounting. Thus, if only a subset of participants in the historical data are exchangeable with the current data, these priors may not be appropriate. In order to combat this issue, propensity score (PS) approaches have been proposed. However, PS approaches are only concerned with the covariate distribution, whereas analysis in practice is typically conducted conditional on the covariates. In this talk, we introduce the latent exchangeability prior (LEAP), where observations in the historical data are classified into exchangeable and non-exchangeable groups. Thus, the LEAP discounts the historical data by identifying the most relevant subjects from the historical data. We compare our proposed approach against alternative approaches in simulations and present a case study using our proposed prior to augment concurrent control in a randomized placebo-controlled Phase 3 clinical trial in plaque psoriasis with historical data.

7 . Incorporating Interventions to an Extended SEIRD Model with Vaccination: Application to COVID-19 in Qatar

[Student Paper Competition 1, (page 5)]

Elizabeth B AMONA, *Virginia Commonwealth University*

The COVID-19 outbreak of 2020 has re-

quired many governments to develop and adopt mathematical-statistical models of the pandemic for policy and planning purposes. To this end, this work provides a tutorial on building a compartmental model using Susceptible, Exposed, Infected, Recovered, Deaths and Vaccinated (SEIRDV) status through time. The proposed model uses interventions to quantify the impact of various government attempts made to slow the spread of the virus. Furthermore, a vaccination parameter is also incorporated in the model, which is inactive until the time the vaccine is deployed. A Bayesian framework is utilized to perform both parameter estimation and prediction. Predictions are made to determine when the peak Active Infections occur. We provide inferential frameworks for assessing the effects of government interventions on the dynamic progression of the pandemic, including the impact of vaccination. The proposed model also allows for quantification of number of excess deaths averted over the study period due to vaccination.

8. Joint Additive Factor Regression for Multi-Omics Data Integration

[01.M1.I6, (page 5)]

Niccolo ANCESCHI, *Duke University*
Federico FERRARI, *Merck & Co., Inc.*
Himel MALLICK, *Cornell University*
David DUNSON, *Duke University*

In precision medicine, it is common to gather data from multiple modalities to characterize different aspects of a patient across biological layers. Such data can lead to more accurate prediction of health responses, motivating principled approaches to integrate modalities. With multi-omics data, signal-to-noise ratio can vary substantially across modalities, which requires more structured statistical tools beyond standard late and early fusion. This challenge comes with the need to preserve interpretability, allowing identification of relevant biomarkers, and proper uncertainty quantification for the predicted outcomes. While these properties are naturally accounted for within a Bayesian framework, state-of-the-art factor analysis (FA) formulations for multi-omics data rely on restrictive modeling assumptions. We propose a novel joint FA model having a structured additive structure, accounting for shared and view-specific components and allowing for flexible covariate and outcome distributions. We provide a fast implementation via MCMC and extend our approach to multi-group settings, to account for interactions among latent factors, and allow longitudinal and spatial data.

9 . Covariance-Based Clustering for Classification

[Poster Session, (page 15)]

Theophilus ANIM BEDIAKO, *South Dakota State University*

Classification involves assigning observations to known classes based on some common features. Fisher's discriminant analysis commonly known as linear and quadratic discriminant analysis(LDA, QDA) are famous classification methods. Both LDA and QDA assume the given classes follow a multivariate normal distribution. In the case of LDA, the classes share a common covariance matrix which could be restrictive. In the case of QDA, each class has a unique covariance matrix, which could be computationally expensive. Clustering methods search for natural groups of similar observations in data. Model-based clustering, particularly mixtures of Gaussian, remains one of the popular clustering methods where the search for classes relies on the assumption that samples within a cluster come from a specific normal distribution. This work proposes a covariance-based clustering method to build a classifier to overcome the simplicity of LDA and the complexity of QDA. For this, a finite mixture of Wishart is developed for clustering the cross-product matrices with similar structures. The parameter estimates of the mixtures are obtained through the Expectation-Maximization(EM) algorithm. Initialization of the parameters for the EM algorithm is proposed. Having identified similar covariance structures in the data, we propose a method, cluster-based LDA, which uses the results of the clustering to build a classifier. The proposed method is studied through simulated data and applied to glass fragment classification

10 . Joint spectral clustering in multilayer networks

[03.A1.I38, (page 24)]

Jesús ARROYO, *Texas A&M University*
Joshua AGTERBERG, *University of Pennsylvania*
Zachary LUBBERTS, *Johns Hopkins University*

Modern network datasets are often composed of multiple layers, such as different views, time-varying observations or independent sample units. These data require flexible and tractable models and methods capable of aggregating information across the

networks. To that end, this talk considers the community detection problem under the multilayer degree-corrected stochastic blockmodel. We propose a spectral clustering algorithm and demonstrate that its misclustering error rate improves exponentially with multiple network realizations, even in the presence of significant layer heterogeneity. The methodology is illustrated in a case study of US airport data, where we identify meaningful community structure and trends influenced by pandemic impacts on travel. This is joint work with Joshua Agterberg and Zachary Lubberts.

11. Kernel Epsilon-greedy strategy for nonparametric bandits

[04.M1.I54, (page 32)]

Sakshi ARYA, *Pennsylvania State University*
Bharath K. SRIPERUMBUDUR, *Pennsylvania State University*

Contextual bandit algorithms are popular for sequential decision-making in several practical applications, ranging from online advertisement recommendations to mobile health. The goal of such problems is to maximize cumulative reward over time for a set of choices/arms while considering covariate (or contextual) information. Epsilon-Greedy is a popular heuristic for the Multi-Armed Bandits problem, however, it is not one of the most studied algorithms theoretically in the presence of contextual information. We study the Epsilon-Greedy strategy in nonparametric bandits, i.e., when no parametric form is assumed for the reward functions. In this work, we assume that the similarities between the covariates and expected rewards can be modeled as arbitrary linear functions of the contexts' images in a specific reproducing kernel Hilbert space (RKHS). We propose a kernelized epsilon-greedy algorithm and establish its convergence rates for estimation and cumulative regret, which are closely tied to the intrinsic dimensionality of the RKHS. We show that the rates closely match the optimal rates for linear contextual bandits when restricted to a finite-dimensional RKHS. Lastly, we illustrate our results through simulation studies and real-data analysis.

12. Discovering underlying dynamics in time series of networks

[01.M2.I9, (page 6)]

Avanti ATHREYA, *Johns Hopkins University*
Zachary LUBBERTS, *Johns Hopkins University*
Youngser PARK, *Johns Hopkins University*
Carey PREIBE, *Johns Hopkins University*

Understanding dramatic changes in the evolution of networks is central to statistical network inference. We consider a joint network model in which each node has an associated time-varying low-dimensional latent vector of feature data, and connection probabilities are functions of these vectors. Under mild assumptions, the time-varying evolution of the constellation of latent vectors exhibits a low-dimensional manifold structure under a suitable notion of distance. This distance can be approximated by a measure of separation between the observed networks themselves, and there exist Euclidean representations for underlying network structure, as characterized by this distance, at any given time. These Euclidean representations and their data-driven estimates permit the visualization of network evolution and transform network inference questions such as change-point and anomaly detection into a classical setting. We illustrate our methodology with real and synthetic data, and identify change points corresponding to shifts in pandemic policies in a communication network of a large organization.

13. Asymptotics of cross validation

[03.A2.I45, (page 27)]

Morgane AUSTERN, *Harvard*
Wenda ZHOU, *NYU/Flatiron*

Cross-validation is a ubiquitous method for risk estimation. However, despite its prevalence, its theoretical properties are still poorly understood. In this talk, we study the asymptotic properties of cross-validation for a large class of models and for an arbitrary number of folds. Under stability conditions, we establish a central limit theorem and Berry-Esseen bounds for the cross-validated risk. This enables us to compute asymptotically consistent confidence intervals for the cross-validated risk and to study the statistical speed-up offered by cross-validation compared to the train-test split procedure. We reveal some surprising behavior of the cross-validated risk and establish the statistically optimal choice for the number of folds.

14. Detection of Structural Breaks in Non-stationary Spatial Random Field

[03.M2.I33, (page 22)]

Pramita BAGCHI, *Department of Statistics, George Mason University*

We propose a method for investigating structural

breaks in a non-stationary spatial random field observed over a regular grid. We work in a frequency domain set-up and propose a statistic based on the maximal difference between local spatial spectral density with maximum taken over locations and range of frequencies. We establish the theoretical properties of this proposed statistic and use that to construct a consistent asymptotic level α test for the stationarity hypothesis. Further, this statistic provides a visual tool to understand the nature of non-stationarity present in the data. We use this visual tool, called disparity map, along with the theoretical properties of this statistic to construct a piece-wise stationary approximation of the observed random field where the pieces are rectangular regions. An initial partition is constructed using a sequential application of the proposed test for stationarity. A hierarchical clustering algorithm is then used to determine the optimal number of regions and to merge the obtained partition appropriately to produce a final partition. In this paper, we present a computationally efficient implementation of our methodology. The accuracy and performance of our proposed methods are demonstrated via extensive simulations and two case studies using climate data.

15 . Temporal Downscaling for Solar Radiation Using a Diurnal Template Model

[Poster Session, (page 15)]

Maggie, D BAILEY, *Colorado School of Mines*

Global and regional climate model projections are useful for gauging future patterns of climate variables, including solar radiation, but data from these models is often too spatio-temporally course for local use. Within the context of solar radiation, the changing climate may have an effect on photo-voltaic (PV) production, especially as the PV industry moves to extend plant lifetimes to 50 years. Predicting PV production while taking into account a changing climate requires data at a resolution that is useful for building PV plants. We present a novel method to downscale global horizontal irradiance (GHI) data from daily averages to hourly profiles, while maintaining spatial correlation of parameters characterizing the diurnal profile of GHI. The method focuses on the use of a diurnal template which can be shifted and scaled according to the time or year and location. Variability in the profile is later added to account for clouds if the daily average value indicates a cloudy day. This analysis is applied to data from the National Solar Radiation Database provided by the National Renewable Energy Lab and a case study of the mentioned methods over California is presented.

16 . High-dimensional Central Limit Theorems for Linear Functionals of Online Least-Squares SGD

[03.E1.I46, (page 28)]

Krishna BALASUBRAMANIAN, *UC Davis*

Stochastic gradient descent (SGD) has emerged as the quintessential method in a data scientist's toolbox. Much progress has been made in the last two decades toward understanding the iteration complexity of SGD (in expectation and high-probability) in the learning theory and optimization literature. However, using SGD for high-stakes applications requires careful quantification of the associated uncertainty. Toward that end, in this work, we establish high-dimensional Central Limit Theorems (CLTs) for linear functionals of online least-squares SGD iterates under a Gaussian design assumption. Our main result shows that a CLT holds even when the dimensionality is of order exponential in the number of iterations of the online SGD, thereby enabling high-dimensional inference with online SGD. Our proof technique involves leveraging Berry-Esseen bounds developed for martingale difference sequences and carefully evaluating the required moment and quadratic variation terms through recent advances in concentration inequalities for product random matrices. We also provide an online approach for estimating the variance appearing in the CLT (required for constructing confidence intervals in practice) and establish consistency results in the high-dimensional setting.

17. Regridding uncertainty for statistical downscaling of solar radiation.

[01.A2.I17, (page 11)]

Soutir BANDYOPADHYAY, *Department of Applied Mathematics and Statistics, Colorado School of Mines*

As the photovoltaic (PV) industry moves to extend plant lifetimes to 50 years, the changing climate may have an effect on PV production and assumptions that current solar radiation patterns are representative of the future may not be appropriate. A key step in aiding the prediction of PV production is projecting solar radiation for future years based on a changing climate. This involves downscaling future climate projections for solar radiation to spatial

and temporal resolutions that are useful for building PV plants. Initial steps in downscaling involve being able to closely predict observed data from regional climate models (RCMs). This prediction requires (1) regridding RCM output from their native grid on differing spatial resolutions to a common grid in order to be comparable to observed data and (2) bias correcting solar radiation data, via quantile mapping, for example, from climate model output. The uncertainty associated with (1) is not always considered for downstream operations in (2). This uncertainty, which is not often shown to the user of a regridded data product, is examined. This analysis is applied to data from the National Solar Radiation Database housed at the National Renewable Energy Lab, and a case study of the mentioned methods in California is presented.

18 . Piecewise and distributed learning using federated methods: statistical considerations and operating characteristics

[01.M1.I1, (page 3)]

Anjishnu BANERJEE, *Medical College of Wisconsin*

Collaborative statistical learning is a paradigm enabling inference from piecewise and distributed data. In particular, we are interested in inference from distributed data preserving privacy of individual data units. Our focus is on ensembles models are trained on separate, private data silos. More specifically, our inferential methods rely on sharing model parameters between units, but never sharing data and maintaing privacy. We describe connections with more general distributed inference and transfer learning settings. We describe general frameworks, when federated joint inference can occur at a central node or by adaptive updation at each node. We investigate several statistical methods that tackle the effect of heterogeneity on the resultant inference and train the data faster and more accurately. We present a novel methods to investigate the effect of heterogeneous variables in such distributed inference procedures on the statistical operating characteristics and consider probability bounds to maintain the validity of inference. We also propose a guarantees for valid inference in these heterogeneous variable settings.

19. Establishing minimax lower bound in high order tensor models for Neuroimaging

[04.M2.I59, (page 34)]

Chitrak BANERJEE, *Wells Fargo Bank, NA*
Lyudmila SAKHANENKO, *Michigan State University*
David C. ZHU, *Michigan State University*

High angular resonance diffusion imaging (HARDI) is a popular in-vivo neuroimaging technique that uses high order tensors to model the anatomy of neural fiber structure inside a live human brain. We develop a minimax bound for the integral curve estimator of the neural fibers from HARDI data. Using this minimax bound, we also develop a novel method to compare different neuroimaging protocols that are commonly used by clinicians. We discuss some interesting simulation results and finally implement this methodology on the HARDI data obtained from a live human brain.

20 . Bayesian inference on Covid-19 transmission dynamics in India using a modified SEIR model

[04.M1.I55, (page 32)]

Paromita BANERJEE, *JOHN CARROLL UNIVERSITY*
Kai YIN, *Case Western Reserve University*
Anirban MONDAL, *Case Western Reserve University*

We propose a modified population-based susceptible-exposed-infectious-recovered (SEIR) compartmental model for a retrospective study of the COVID-19 transmission dynamics in India during the first wave. We extend the conventional SEIR methodology to account for the complexities of COVID-19 infection, its multiple symptoms, and transmission pathways. In particular, we consider a time-dependent transmission rate to account for governmental controls (e.g., national lockdown) and individual behavioral factors (e.g., social distancing, mask-wearing, personal hygiene, and self-quarantine). An essential feature of COVID-19 that is different from other infections is the significant contribution of asymptomatic and pre-symptomatic cases to the transmission cycle. A Bayesian method is used to calibrate the proposed SEIR model using publicly available data (daily new tested positive, death, and recovery cases) from several Indian states. The uncertainty of the parameters is naturally expressed as the posterior probability distribution. The calibrated model is used to estimate undetected cases and study different initial intervention policies, screening rates, and public behavior factors, that can potentially strike a balance between disease control

and the humanitarian crisis caused by a sudden strict lockdown.

21. Attributed random networks: local weak limits, PageRank and sampling fairness

[03.A1.I35, (page 24)]

Sayan BANERJEE, *University of North Carolina, Chapel Hill*
Nelson ANTUNES, *Center for Computational and Stochastic Mathematics,University of Lisbon*
Shankar BHAMIDI, *UNC Chapel Hill*
Vladas PIPIRAS, *UNC Chapel Hill*

We will analyze growing random networks, comprising nodes with different attribute types (communities), which evolve under the combined effect of node popularity (graph degree) and interaction between attributes. Using stochastic approximation techniques, we will show that local neighborhoods of typical nodes in large networks approach limiting random graphs given in terms of randomly stopped multi-type branching processes. This gives detailed information on the limiting distributions of centrality scores like degree and PageRank. We also see how our theoretical results shed light on sampling schemes ensuring prescribed representation of minorities, and on competition between majority and minority communities in the growing network. Based on joint work with Nelson Antunes, Shankar Bhamidi and Vladas Pipiras.

22. Bayesian cooperative learning with BART

[01.M1.I6, (page 5)]

Piyali BASAK, *Merck & Co.*
Himel MALLICK, *Cornell University*
Erina PAUL, *Merck & Co.*

We propose a new Bayesian method for multi-omics data integration. We formulate the multi-omics integration problem as a Bayesian cooperative learning problem that uses an "agreement" shrinkage to encourage agreement of predictions from different data layers. Our method combines Bayesian additive regression trees (BART) with the agreement shrinkage and we show that our Bayesian cooperative learning method achieves higher predictive accuracy on multi-omics datasets, outperforming competing methods while also facilitating uncertainty quantification.

23. Joint modeling with high dimensional longitudinal processes

[04.M1.I51, (page 31)]

Sanjib BASU, *UIC*

TBA

24. Modeling High-Dimensional Spatial Covariance Structures for Climate Processes using Physically-Informed Basis Functions

[03.A1.I37, (page 24)]

Samuel BAUGH, *Lawrence Berkeley National Laboratory*
Samuel BAUGH, *Lawrence Berkeley National Lab*
Karen MCKINNON, *UCLA*
Mark RISSER, *Lawrence Berkeley National Lab*

Accurately quantifying the variability of the global climate system under natural and anthropogenic scenarios is essential for producing reliable assessments of climate change. Large ensembles of general circulation models, run with initial condition perturbations, are invaluable resources for estimating the global structure of variability. However, the high dimension of the climate process covariance matrix relative to the number of ensemble members requires imposed structure in the parameterization to achieve a reliable estimate of variability. In this work, we propose physically-informed basis functions to flexibly model the covariance structure of a climate process while avoiding bias from over-fitting. These basis functions permit a more accurate representation of the underlying covariance structure compared to commonly-used Gaussian process and principal component-based approaches, while maintaining significant computational advantages. These benefits are demonstrated through a climate change "detection and attribution" case study, where the use of these basis functions within a hierarchical Bayesian model enables the feasible representation of multiple complex sources of uncertainty in determining the degree of human responsibility for observed climate change.

25. Performance guarantees of Spectral clustering using g-distance and Longest leg path distance

[Poster Session, (page 15)]

Sabyasachi BERA, *University of Minnesota*

We consider the problem of spectral clustering (SC) with g-distance and the longest-leg path dis-

tance (LLPD) metric, which are able to capture both the geometry and the density information from the data unlike Euclidean distance. We show that SC with g-distance is an interpolation between traditional centroid based clustering and mode based clustering. We prove guarantees on the performance of SC with respect to these metrics when random samples are drawn from different data generating models such as a d dimensional Riemannian manifold for d $>$ 1, multiple lower dimensional Riemannian manifolds embedded inside ambient high dimensional space with or without the presence of background noise, multiple self-intersecting Riemannian manifolds etc. In particular, we provide conditions under which the Laplacian eigengap statistic correctly determines the number of clusters under these data generating models and prove guarantees on the labeling accuracy of the proposed algorithms while simultaneously comparing the performance of different g distances and LLPD.

26. Maximum likelihood estimation of response propensity to a nonprobability survey

[01.A2.I18, (page 11)]

Vladislav BERESOVSKY, *Bureau of Labor Statistics*
Terrance SAVITSKY, *BLS*
Matthew WILLIAMS, *RTI*
Julie GERSHUNSKAYA, *BLS*

Non-probability, or convenience, surveys attract significant attention as a less expensive and more expedient alternative to probability surveys for estimating population characteristics. For design-based inferences a convenience sample may be treated as selected by Poisson sampling from the population of interest. While response indicator to a non-probability survey is unobserved, response propensity *pi_c* may still be estimated by maximizing Bernoulli likelihood of the observed indicator Z of the convenience sample integrated with a probability sample from the same population. This is possible by utilizing the direct relationship between *pi_ci*, Bernoulli parameter *pi_zi* and random sampling probability *pi_ri* for the units i of the integrated sample. We demonstrate that the proposed method is more efficient than maximizing a pseudo-likelihood by solving generalized estimating equations. We proceed to investigate dependence of the stability of the estimated response propensities on the sizes and extent of population coverages by the non-probability and probability samples.

27. A Multivariate Bayesian Hierarchical Model for Small Area Estimation of Criminal Victimization Rates in Domains Defined by Age and Gender

[04.M2.I58, (page 33)]

Emily BERG, *Iowa State University*

The National Crime Victimization Survey (NCVS) gathers information on criminal victimizations for individuals in a representative sample of United States households. The NCVS provides authoritative data on the rates of many types of violent crimes, including simple assault, robbery, and aggravated assault. The NCVS is designed to permit direct estimation for 22 large states and for the nation. Small sample sizes preclude production of direct estimates for more detailed domains. Past work on small area estimation with the NCVS data has focused heavily on geographic subdivisions of the United States. Estimates for demographic subdivisions are also of interest. We consider the problem of producing estimates for small domains defined by the intersection of age categories and genders. We construct estimates for four types of violent crimes in each of two time periods. We accomplish this through the use of a multivariate Bayesian hierarchical model. We compare a model with a log transformation to a model fit to the data in the original scale. We compare small area predictors based on a selected model to the direct estimators.

28. Individualized Inference in Bayesian Quantile Directed Acyclic Graphical Models

[01.E1.I21, (page 13)]

Anindya BHADRA, *Purdue University*
Ksheera SAGAR, *Purdue University*
Yang NI, *Texas A&M University*
Veera BALADANDAYUTHAPANI, *University of Michigan*

We propose an approach termed "qDAGx" for Bayesian covariate-dependent quantile directed acyclic graphs (DAGs) where these DAGs are individualized, in the sense that they depend on individual-specific covariates. A key distinguishing feature of the proposed approach is that the individualized DAG structure can be uniquely identified at any given quantile, based on purely observational data without strong assumptions such as a known topological ordering. For scaling the proposed method to a large number of variables and covariates, we

propose for the model parameters a novel parameter expanded horseshoe prior that affords a number of attractive theoretical and computational benefits to our approach. By modeling the conditional quantiles, qDAGx overcomes the common limitations of mean regression for DAGs, which can be sensitive to the choice of likelihood, e.g., an assumption of multivariate normality, as well as to the choice of priors. We demonstrate the performance of qDAGx through extensive numerical simulations and via an application in precision medicine by inferring patient-specific protein–protein interaction networks in lung cancer.

29. Scalable Community Detection in Massive Networks via Predictive Assignment

[Poster Session, (page 15)]

Subhankar BHADRA, *North Carolina State University*

Massive network data are becoming increasingly common in scientific applications. Existing community detection methods are computationally infeasible for such massive networks due to two reasons. First, the full network needs to be stored and analyzed in a single server, leading to exorbitant memory costs. Second, existing methods typically use matrix factorization or iterative optimization using the full network, leading to impractically high runtimes. We propose a strategy called predictive assignment to enable computationally efficient community detection while ensuring statistical accuracy. The core idea is to avoid large-scale matrix computations by breaking up the task into a smaller matrix computation plus a large number of vector computations which can be carried out in parallel. Under the proposed method, community detection is carried out on a small subgraph to estimate the relevant model parameters. The next step is predictive assignment, a new approach where we assign each remaining node to a community based on these estimates. We study the theory and the practice of predictive assignment for spectral clustering and bias-adjusted spectral clustering under the stochastic blockmodel and its degree-corrected version. We also illustrate the benefits of this new method on two large real-world datasets: the Digital Bibliography & Library Project (DBLP), a computer science bibliographical database, and the Twitch Gamers Social Network.

30. Estimating Prediction Error for Functional Time Series

[03.M2.C1, (page 23)]

Samayita BHATTACHARJEE, *University of California, Davis*
Alexander AUE, *University of California, Davis*
Prabir BURMAN, *University of California, Davis*

With the emergence of modern technology and availability of high frequency data, study of functional time series is becoming popular in recent years. One of the main goals of time series is to predict the future. Though some methods of predicting a functional time series have been explored in the literature, not much work has been done to get the estimates of prediction error. This research aims to answer that question by proposing estimates of prediction error in the functional time series context. Prediction error is useful to determine how good the underlying model fits the data and hence estimating it is important. Prediction error is also important for constructing prediction bands and this paper also aims to address that question by proposing methods of getting prediction bands using the different functional prediction error estimates. The research shows results based on simulation study as well as real data applications.

31. On the Convergence of Coordinate Ascent Variational Inference

[03.A1.I40, (page 25)]

Anirban BHATTACHARYA, *Texas A&M University*
Debdeep PATI, *Texas A&M University*
Yun YANG, *University of Illinois at Urbana-Champaign*

As a computational alternative to Markov chain Monte Carlo approaches, variational inference (VI) is becoming increasingly popular for approximating intractable posterior distributions in large-scale Bayesian models due to its comparable efficacy and superior efficiency. Several recent works provide theoretical justifications of VI by proving its statistical optimality for parameter estimation under various settings; meanwhile, formal analysis on the algorithmic convergence aspects of VI is still largely lacking. In this talk, we will discuss some recent advances towards studying convergence of the popular coordinate ascent variational inference algorithm. We will present some specific case studies and proceed to develop a general framework for studying such questions.

32. Estimation of the number of communities for sparse networks

[03.E1.I50, (page 29)]

Sharmodeep BHATTACHARYYA, *Oregon State University*
Neil HWANG, *Bronx Community College, City University of New York*
Jiarui (Sam) XU, *Meta Inc.*
Shirshendu CHATTERJEE, *City College, City University of New York*

Among the non-parametric methods of estimating the number of communities (K) in a community detection problem, methods based on the spectrum of the Bethe Hessian matrices (H_ζ with the scalar parameter ζ) have garnered much popularity for their simplicity, computational efficiency, and robustness to the sparsity of data. For certain heuristic choices of ζ, such methods have been shown to be consistent for networks with N nodes with a common expected degree of $\omega(\log N)$. In this paper, we obtain several finite sample results to show that if the input network is generated from either stochastic block models or degree-corrected block models, and if ζ is chosen from a certain interval, then the associated spectral methods based on H_ζ is consistent for estimating K for the sub-logarithmic sparse regime, when the expected maximum degree is $o(\log N)$ and $\omega(1)$, under some mild conditions even in the situation when K increases with N. We also propose a method to empirically estimate the aforementioned interval, enabling us to develop a consistent K estimation procedure in the sparse regime. We evaluate the performance of the resulting estimation procedure's performance theoretically and empirically through extensive simulation studies and application to a comprehensive collection of real-world network data.

33. Structured Dynamic Pricing: Optimal Regret in a Global Shrinkage Model

[Poster Session, (page 15)]

Rashmi Ranjan BHUYAN, *University of Southern California*

We consider dynamic pricing strategies in a streamed longitudinal data set-up where the objective is to maximize, over time, the cumulative profit across a large number of customer segments. We consider a dynamic probit model with the consumers' preferences as well as price sensitivity varying over time. Building on the well-known finding that consumers sharing similar characteristics act in similar ways, we consider a global shrinkage structure, which assumes that the consumers' preferences across the different segments can be well approximated by a spatial autoregressive (SAR) model. In such a streamed longitudinal set-up, we measure the performance of a dynamic pricing policy via regret, which is the expected revenue loss compared to a clairvoyant that knows the sequence of model parameters in advance. We propose a pricing policy based on penalized stochastic gradient descent (PSGD) and explicitly characterize its regret as functions of time, the temporal variability in the model parameters as well as the strength of the auto-correlation network structure spanning the varied customer segments. Our regret analysis results not only demonstrate asymptotic optimality of the proposed policy but also show that for policy planning it is essential to incorporate available structural information as policies based on unshrunken models are highly sub-optimal in the aforementioned set-up.

34. A prescreening methodology for the use of likelihood ratios with subpopulation structures in the alternative source population

[Poster Session, (page 15)]

Dylan, D BORCHERT, *South Dakota State University*

Prescreening is a commonly used methodology in which the forensic examiner builds a background population or a relevant source population with sources or objects that meet a prespecified degree of similarity to the given piece of evidence. This relevant source population is then used to give a value of evidence in the form of a likelihood ratio or a Bayes factor. An advantage of prescreening is it can protect from a known or unknown subpopulation structure in the alternative source population by isolating the subpopulation from which the given evidence has arisen. This paper discusses a prescreening methodology used in conjunction with two commonly used likelihood ratios, as well as a score-based analogue. An extensive simulation study with synthetic and real data suggested models was conducted along with real trace element data examples. We find that prescreening can help give an accurate value of evidence when there is a subpopulation structure in the alternative source population, but it can also give a more extreme value than the true value of evidence within the subpopulation of interest in certain scenarios. The results suggest that prescreening can be useful to present a

value of evidence relative to the subpopulation of interest. The authors suggest that if prescreening is implemented prior to the evaluation of a value of evidence in the form of a likelihood ratio, the prescreening level should be reported alongside the value of evidence.

35. Limiting spectral distribution of random matrices with independent entries

[Memorial Session 3, (page 28)]

Arup BOSE, *Indian Statistical Institute*
Priyanka SEN, *Indian Statistical Institute*
Arusharka SEN, *Concordia University*
Koushik SAHA, *koushiksaha877@gmail.com*

It is well known that the limit eigenvalue distribution of the scaled standard Wigner matrix is the semi-circular distribution whose $2k$th moment equals the number of non-crossing pair-partitions of $\{1, 2, \ldots, 2k\}$. There are several extensions of this result in the literature, including the sparse case. We discuss extension of these results by relaxing significantly the i.i.d. assumption. The limiting spectral distribution then involve a larger class of partitions. In the process we show how some new sets of partitions gain importance. Several existing and new results for their band and sparse versions, as well as for matrices with continuous and discrete variance profile follow as special cases.

Patterned random matrices such as the reverse circulant, the symmetric circulant, the Toeplitz and the Hankel matrices and their almost sure limiting spectral distribution (LSD), have also been studied quite extensively. Under the assumption that the entries are taken from an i.i.d. sequence with finite variance, the LSD are tied together by a common thread –the $2k$th moment of the limit equals a weighted sum over different types of pair-partitions of the set $\{1, 2, \ldots, 2k\}$ and are universal. Some results are also known for the sparse case. Time permitting we show suitable extension of these results also, along the lines of the Wigner matrix.

36 . Deep Hierarchical Generalized Transformation Models with Application to Wildfires

[03.E1.I49, (page 29)]

Jonathan BRADLEY, *Florida State University*

In recent years, wildfires have devastated communities and represent an immediate danger to humans in terms of property damage and death. There is evidence that wildfires are becoming more frequent and are increasing in size. Moreover, wildfires are predicted to continue to increase in a warming climate. Prediction of wildfires can be difficult when one acknowledges that firebrand showers are known to be a complex nonlinear stochastic process. Motivated by an extension of the Hierarchical Generalized Transformation (HGT) model, we develop a new class of Bayesian neural network models to analyze unknown nonlinear functions that treat the activation functions as unknown. Traditional BNNs have considerably more computational difficulties than standard neural network models that make use of efficient backpropagation algorithms. We impose conditional independence assumptions among the activation functions that lead to an efficient implementation of our BNN. This particular type of BNN can be interpreted as a nested or "deep" hierarchical generalized transformation model (DHGT). An analysis of a recent California wildfire using a DHGT is presented.

37. Combining Data Sources to Produce Nationally Representative Estimates of Hospital Encounter Characteristics

[01.M1.I3, (page 3)]

Jay BREIDT, *NORC at the University of Chicago*
Dean RESNICK, *NORC at the University of Chicago*
Geoffrey JACKSON, *National Center for Health Statistics*
Donielle WHITE, *National Center for Health Statistics*

The 2020 National Hospital Care Survey (NHCS) is a stratified random sample of US hospitals, conducted by the Centers for Disease Control and Prevention's National Center for Health Statistics (NCHS). Hospitals responding to NHCS provide nearly complete records of patient encounters over the entire 2020 calendar year, making the data extraordinarily valuable for understanding US hospital care utilization and informing health care policy. NHCS is subject to hospital-level nonresponse that reduces available sample sizes and potentially biases results, due to differential response rates across hospital types. Accordingly, NCHS and NORC at the University of Chicago have collaborated to enhance NHCS encounter data with additional data sources, reducing potential biases. The additional data sources include a proprietary commercial hospital encounter data source, treated as a nonprobability sample with unknown hospital participation

propensities, as well as nationally representative hospital care benchmarks from the Healthcare Cost and Utilization Project (HCUP). The enhanced data can be used to create nationally representative estimates of hospital encounter characteristics. The enhanced data will also serve as the basis for additional data products, including a weighted public use file and experimental synthetic data products. We will describe our data enhancement approach along with methodological challenges and preliminary results.

38 . Bringing statistical innovation into pharmaceutical drug development: Closed MCP-Mod for pairwise comparisons of several doses with a control

[Special Invited Session 4, (page 18)]

Frank BRETZ, *Novartis*

Statistical methodologies in pharmaceutical drug development have evolved dramatically over the past decades in response to a constantly evolving environment. In this presentation we discuss how advances in technology drive the need for new statistical approaches and illustrate these cycles of innovation with a concrete example. In the first part we employ the themes and language commonly used in drug development to describe how statistical methods and tools have evolved over time, covering the three great themes of statistical inference, statistical modelling and statistical learning. In the second part we then illustrate this evolution by describing an extension of the MCP-Mod dose finding approach to confirmatory clinical trials, thereby bridging classical statistical inference with statistical modelling. The MCP-Mod approach is commonly applied to dose response testing and estimation in exploratory clinical trials. The MCP part of MCP-Mod was originally developed to detect a dose response signal using a multiple contrast test, but it is not appropriate to make a specific claim that the drug has a positive effect at an individual dose. In this presentation we extend the MCP-Mod approach to obtain confirmatory p-values for detecting a dose response signal as well as for the pairwise comparisons of the individual doses against placebo.

39. Frequency Band Analysis of Nonstationary Multivariate Time Series

[03.M2.I33, (page 22)]

Scott BRUCE, *Texas A&M University*
Raanju SUNDARARAJAN, *Southern Methodist University*
Scott BRUCE, *Texas A&M University*

Information from frequency bands in biomedical time series provides useful summaries of the observed signal. Many existing methods consider summaries of the time series obtained over a few well-known, pre-defined frequency bands of interest. However, these methods do not provide data-driven methods for identifying frequency bands that optimally summarize frequency-domain information in the time series. A new method to identify partition points in the frequency space of a multivariate locally stationary time series is proposed. These partition points signify changes across frequencies in the time-varying behavior of the signal and provide frequency band summary measures that best preserve the nonstationary dynamics of the observed series. An L_2 norm-based discrepancy measure that finds differences in the time-varying spectral density matrix is constructed, and its asymptotic properties are derived. New nonparametric bootstrap tests are also provided to identify significant frequency partition points and to identify components and cross-components of the spectral matrix exhibiting changes over frequencies. Finite-sample performance of the proposed method is illustrated via simulations. The proposed method is used to develop optimal frequency band summary measures for characterizing time-varying behavior in resting-state electroencephalography (EEG) time series, as well as identifying components and cross-components associated with each frequency partition point.

40 . Bayesian methods for vaccine safety surveillance using federated data sources

[01.M1.I1, (page 3)]

Fan BU, *UCLA*

We discuss a Bayesian sequential analysis framework for data sources that are distributed across a federated network, motivated by vaccine safety surveillance studies. We wish to enable rapid detection of vaccine safety events from observational healthcare data that accrue over time. Our framework aims at resolving three main challenges: first, control of testing errors in sequential analyses of streaming data; second, correction of bias induced by observational data; third, distributed learning of federated data sources while preserving patient-level privacy. Through extraction of profile likelihoods

that retain rich distributional information while protecting individual-level data privacy, and hierarchical analysis of negative control outcomes, we tackle these challenges in a unified statistical framework. As evidenced by large-scale empirical evaluations using real-world data sources, our framework provides substantial improvements over existing approaches of safety surveillance.

41 . Large Deviations and Stochastic Control

[Bahadur Memorial Lecture, (page 17)]

Amarjit BUDHIRAJA, *University of North Carolina Chapel Hill*

The theory of large deviations has a long history with deep connections to statistical mechanics, information theory, partial differential equations, optimization theory, and mathematical statistics. In particular, Professor Bahadurâ€™s seminal work in asymptotic statistics has laid key mathematical foundations in the field. In recent years an approach for analyzing large deviation asymptotics for stochastic dynamical systems has emerged, that transforms the study of large deviation probabilities to that of the asymptotic behavior of certain stochastic control problems. One of the key advantages of this approach is that, instead of needing to establish exponential probability estimates, the main ingredient is proving suitable tightness and weak convergence properties. In this talk I will give a high-level overview of this approach and provide a survey of some of the large deviation problems that have been studied using this method.

42. A general framework for regression with mismatched data

[Poster Session, (page 15)]

Priyanjali BUKKE, *George Mason University*

Data analysis is often based on files that are the result of merging and integrating multiple data sets from different sources. In data integration, record linkage is an essential task for linking records across data sets that refer to the same entity. Record linkage is not error-free; there is a possibility that records belonging to different entities are mismatched, or that records belonging to the same entity are not identified. As a result, linkage error can significantly reduce the quality of the resulting data. In subsequent statistical analyses, it is, therefore, advisable to make suitable adjustments that account for potential bias caused by data contamination or sample selection introduced by record linkage. In this poster, we present a general framework for regression with mismatched data to help enable reliable post-linkage data analysis and inference.

43 . Statistical and Ethical Considerations in the Analysis of Mobile Phone Tracking Data

[Special Invited Session 1, (page 8)]

Catherine CALDER, *University of Texas at Austin*
Marcin JUREK, *University of Texas at Austin*
Corwin ZIGLER, *University of Texas at Austin*

In the biomedical and social sciences, mobile phone tracking (MPT) data — collected using location sensing technologies readily available on smartphones — has become an increasingly common component of cohort studies, where it has been employed for purposes of digital phenotyping or estimating personal exposure to the ambient environment or particular social contexts. In this talk, I will provide an overview of some of the statistical challenges that arise when working with MPT data for research purposes. I will then provide an in depth investigation into the consequences of MPT study design choices from a formal missing data perspective. To do so, I will introduce a novel statistical model that formalizes the so-called flight-pause paradigm for human movement as a likelihood for a random object, called a motion, made up of increments of changes in space and time. Under this model, it is possible to perform both inference on unknown model parameters and trajectory imputation under various forms of missing data that are ubiquitous in practice. Under this model, it is possible to illuminate the consequences of different MPT data collection mechanisms, including the surprising result that common assumptions about the missing data mechanism for MPT are not valid for the mechanism governing the random motions of the flight-pause model. The consequences of missing data and proposed adjustments will be illustrated using both simulations and real data, illustrating how the statistical formulation pursued here can serve as a foundation for continued statistical research on MPT data collection, design, and analysis. Finally, I will briefly discuss some ethical considerations related to the use of MPT data for research purposes.

44. Variational sparse inverse Cholesky approximation for latent Gaussian pro-

cesses via double Kullback-Leibler minimization

[01.M2.I10, (page 7)]

Jian CAO, *Texas A&M University*
Myeongjong KANG, *Texas A&M*
Felix JIMENEZ, *Texas A&M*
Matthias KATZFUSS, *Texas A&M*

To achieve scalable and accurate inference for latent Gaussian processes, we propose a variational approximation based on a family of Gaussian distributions whose covariance matrices have sparse inverse Cholesky (SIC) factors. We combine this variational approximation of the posterior with a similar and efficient SIC-restricted Kullback-Leibler-optimal approximation of the prior. We then focus on a particular SIC ordering and nearest-neighbor-based sparsity pattern resulting in highly accurate prior and posterior approximations. For this setting, our variational approximation can be computed via stochastic gradient descent in polylogarithmic time per iteration. We provide numerical comparisons showing that the proposed double-Kullback-Leibler-optimal Gaussian-process approximation (DKLGP) can sometimes be vastly more accurate than alternative approaches such as inducing-point and mean-field approximations at similar computational complexity.

45. On network modularity statistics in connectomics and schizophrenia

[03.A1.I35, (page 23)]

Joshua CAPE, *University of Wisconsin*
Anirban MITRA, *University of Pittsburgh*
Konasale PRASAD, *University of Pittsburgh*

Modularity-based methods for structure and community discovery remain popular in the network neuroscience literature and enjoy a history of yielding meaningful neurobiological findings. All the while, the full potential of these methods remains limited in part by an absence of uncertainty quantification guarantees for use in downstream statistical inference. Here, we pursue this direction by revisiting the classical notion of modularity maximization in the analysis of adjacency and correlation matrices. We begin by considering certain latent space network models wherein high-dimensional matrix spectral properties can be precisely analyzed. We further propose and argue for the potential usefulness of several new, non-classical modularity-type network statistics. Our findings are applied to an analysis of dMRI and fMRI data in the study of schizophrenia.

46. Debiasing in missing data models with inaccurate nuisance parameters

[03.E1.I46, (page 28)]

Michael CELENTANO, *UC Berkeley*

We consider of (i) estimating linear model coefficients with data missing at random (MAR) and (ii) average treatment effect estimation with linear outcome models under strong ignorability in a high-dimensional regime in which the number of confounders is proportional to the sample size and the outcome and propensity/missingness models cannot be estimated consistently. In the case $n > p$, Jiang et al. 2022 and Yadlowksy 2022 developed theory for the classical AIPW estimator in this regime and established its variance inflation and asymptotic normality when the outcome model is fit by ordinary least squares. In this paper, we study the case $n < p$, where ordinary least squares is not feasible. We show that classical estimators of linear effects and average treatment effects can be biased and inconsistent, and standard debiasing adjustments relying on oracle inverse propensity weights fail to give unbiased estimates of the linear effects. We propose an alternative estimator that is provably consistent for the average treatment effect, and we provide confidence intervals for the linear model coefficients. Our proposed estimator requires estimation of both the outcome and propensity/missingness models, but we combine these estimates in a non-standard way.

47. Subsampling Based Community Detection for Large Networks

[Poster Session, (page 16)]

Sayan CHAKRABARTY, *University of Illinois at Urbana Champaign*

Large networks are becoming pervasive in scientific applications. Statistical analysis of such large networks is prohibitive due to exorbitant runtime and high memory requirements. We propose a subsampling based divide-and-conquer algorithm, SONNET, for community detection in large networks. The algorithm splits the original network into multiple subnetworks with a common overlap, and carries out detection algorithm for each subnetwork. The results from individual subnetworks are aggregated using a label matching method to get the final community labels. This method saves both mem-

ory and computation costs significantly as one needs to store and process only the smaller subnetworks. This method is also parallelizable which makes it even faster.

48. Robust probabilistic inference via a constrained transport metric

[Student Paper Competition 1, (page 5)]

Abhisek CHAKRABORTY, *Texas A&M University*

Flexible Bayesian models are typically constructed using limits of large parametric models with a multitude of parameters that are often uninterpretable. In this article, we oer a novel alternative by constructing an exponentially tilted empirical likelihood carefully de- signed to concentrate near a parametric family of distributions of choice with respect to a novel variant of the Wasserstein metric, which is then combined with a prior distribution on model parameters to obtain a robustied posterior. The proposed approach finds applications in a wide variety of robust inference problems, where we intend to perform inference on the parameters associated with the centering distribution in presence of outliers. Our proposed transport metric enjoys great computational simplicity, exploiting the Sinkhorn regularization for discrete optimal transport problems, and being inherently parallelizable. We demonstrate superior performance of our methodology when compared against state-of-the-art robust Bayesian inference methods. We also demonstrate equivalence of our approach with a nonparametric Bayesian formulation under a suitable asymptotic framework, testifying to its flexibility. The constrained entropy maximization that sits at the heart of our likelihood formulation finds its utility beyond robust Bayesian inference; an illustration is provided in a trustworthy machine learning application.

49. A bootstrap test for MANOVA in high dimensions

[01.A1.I12, (page 8)]

Nilanjan CHAKRABORTY, *Washington University in Saint Louis*
Lyudmila SAKHANENKO, *Michigan State University*

This talk mainly concerns about a K sample High dimensional CLT over a class of Hyper-Polygons. This result finds an extremely useful application towards formulating a bootstrap based test for High Dimensional MANOVA problem based on supremum type test statistics for the difference in means among the K groups. The test procedure considered is free from any distribution and correlational assumptions which broadens its scope towards practical applications. The problem of Generalized Linear Hypothesis testing problem has also been tackled using the previously mentioned CLT result. The asymptotic analysis of these tests in terms of controlling size and power has been theoretically validated. A detailed simulation study has been conducted which corroborates the findings done in the theoretical section.

50. Scalable Bayesian Variable Selection and Grouping for Binary and Multiclass Outcome Data

[04.M1.I55, (page 33)]

Sounak CHAKRABORTY, *University of Missouri*

TBA

51. On the Analysis of Large Scale Observational Streaming Data

[Poster Session, (page 16)]

Aleena CHANDA, *University of Nebraska-Lincoln*

The main goal of this paper is to make one step ahead predictions in an M-open streaming data context. We partition the range of data into intervals, observe the data until a specified time t, and form a histogram from which we obtain a prediction. To ensure a running time bound we reduce the data stream to a representative set via the doubling algorithm and sequential k-means at each time step before forming our predictor. Since, we do not assume any distribution of the data, we introduce randomness via. hash functions.

We verify that our proposed method satisfies an error bound and converges in probability and almost surely(a.s.) to a recognizable limit. Then we compare our method to predictors based on Gaussian Process Priors and Dirichlet Process Priors. We find that our method performs the best or ties for best among the three methods that we have implemented so far.

52. How to compare two curves: the easiest question in survival analysis?

[01.M2.I8, (page 6)]

Rick CHAPPELL, *University of Wisconsin*
Mitchell PAUKNER, *University of Wisconsin*

How to compare two curves: the easiest question in survival analysis?

Rick Chappell Professor, Depts. of Statistics and of Biostatistics and Medical Informatics University of Wisconsin Madison

Textbooks describing how to analyze time-to-event outcomes in clinical trials tend to list a limited range of topics. Differences are often quantified using hazard ratios from the Cox model and its associated score, the log-rank test. Weighted rank tests may be presented, along with comparisons of landmarks and quantiles. All these have their disadvantages in terms of interpretation, convenience, and/or power. Furthermore, cancer immunotherapy and many other treatments in a variety of diseases can have delayed effects causing pairs of curves to diverge after months or years of followup. In such cases the log-rank test's power will be low and the associated hazard ratio estimate uninterpretable. Rank tests with increasing weights suffer from the paradoxical property of rewarding early failures. I will discuss various available alternatives including some which are quite new.

53. What is the SCT and what's in it for you - a historical journey

[01.E1.I22, (page 13)]

Rick CHAPPELL, *University of Wisconsin*

TBA

54. Flexible Tree-based Model for Extreme Events

[03.M2.C1, (page 23)]

Suneel Babu CHATLA, *University of Texas at El Paso, Texas*
Galit SHMUELI, *National Tsing Hua University, Hsinchu, Taiwan*

Tree-based methods offer a flexible approach to modeling data while also being known to provide higher levels of interpretability. Modeling extreme or rare events is useful in a variety of applications in fields ranging from meteorology to economics. We present a flexible tree-based approach for modeling extreme values. Our approach considers the two prominent paradigms to model extreme events: block-maxima and peak-over-threshold. For estimation of the tree, we use a model-based recursive partitioning algorithm which uses coefficient constancy tests for identifying the split variable and uses an exhaustive search for estimating the split point at each node. In addition, the proposed approach also allows to include a fixed or global effect which do not vary across the nodes in the fitted tree. In general, to provide a greater flexibility in modeling, the fixed effects are considered as nonparametric effects. We illustrate the usefulness of the proposed methodology using bikesharing data from San Francisco bay area.

55. An Agnostic Fay-Herriot Model For Small Area Statistics

[04.M1.I52, (page 31)]

Ansu CHATTERJEE, *University of Minnesota*

We propose an extension of the Fay-Herriot model to include cases where an underlying distribution in the hierarchical structure may be non-Gaussian. A Gaussian process-based Bayesian technique is developed for this extended framework. We compare the performance of the traditional Gaussianity-based empirical best linear unbiased predictor (EBLUP) and a hierarchical Bayesian prediction technique with the proposed methodology. It is observed that while Bayesian predictors and some frequentist alternatives perform well in some circumstances, the proposed extension fo the Fay-Herriot method is more accurate when Gaussianity is suspect, thus lending robustness to small area studies.

56. A Theoretically Tractable Cross Validation Framework for Signal Denoising

[03.A2.I45, (page 27)]

Sabyasachi CHATTERJEE, *University of Illinois at Urbana Champaign*

TBA

57. Change point detection, estimation, and localization for network data

[03.E1.I50, (page 29)]

Shirshendu CHATTERJEE, *City University of New York*
Soumendu Sundar MUKHERJEE, *Indian Statistical Institute*
Sharmodeep BHATTACHARYYA, *Oregon State University*

We will discuss the offline change point detection and localization problem in the context of piecewise stationary network data, where the observable is a finite sequence of networks. We will discuss the associated challenges, detectability and localizability thresholds, available algorithms, and their comparisons.

58. CMPLE to decode Photosynthesis using MM algorithm

[03.M2.C1, (page 23)]

Abhijnan CHATTOPADHYAY, *Postdoctoral Fellow, National Institute of Environmental Health Science, National Institute of Health*
David Mark KRAMER, *Professor, Plant Research Lab, Michigan State University*
Tapabrata MAITI, *Professor, Department of Statistics, Michigan State University*
Samiran SINHA, *Professor, Department of Statistics, Texas A & M University*

In quantitative genomic experiments, correlations among various biological responses (phenotypes) give new insights into how genetic diversity may have tuned biological processes to enhance fitness under diverse conditions. However, the current literature lacks any method for assessing the effect of predictors on pairwise correlations among multiple phenotypes together with easily interpretable model parameters. To address this need, we propose to directly model pairwise correlations in terms of genetic and environmental variables and develop a computationally efficient inference procedure. There are two major novelties in our methodology: firstly, we use a composite pairwise likelihood method to avoid definiteness restrictions on the correlation matrix, and secondly, we develop a novel Minorize-Maximize (MM) algorithm for the efficient estimation of a large number of parameters. Numerical experiment on synthetic data shows excellent performance of the proposed method in terms of characteristics of the estimators and computational efficiency. Analyzing a dataset from recombinant inbred cowpea lines, the method helps distinguish mechanisms by which genetic components impact a common aggregate phenotype. Specifically, we show that the rates of solar energy storage by photosynthesis (the aggregate trait) are differentially affected by different genetic loci through two distinct processes: "photoinhibition," which results from photodamage caused by excess light, and "photoprotection," which protects plants from photodamage but also results in energy loss.

59. On an Empirical likelihood-based Estimator for Complex Survey Data

[03.A2.I44, (page 27)]

Sanjay CHAUDHURI, *University of Nebraska-Lincoln*

The empirical likelihood-based methods provide interesting ways to analyze complex survey data. Using various estimating equations, it easily incorporates many model and population-level constraints. It can also be used to justify a semi-parametric likelihood-based inference of model parameters when due to the complex sampling procedure the observed data has a distribution different from the model. We discuss the implications of a specific form of constraint on empirical likelihood-based parameter estimates. We focus on the properties of the optimal weights as well as on the estimation of standard errors.

60. On a modified deep neural network based mass imputation for data integration

[01.M1.I3, (page 4)]

Sixia CHEN, *University of Oklahoma Health Sciences Center*

TBA

61. Evaluation of small-area estimators and associated MSPE estimators under model misspecification

[Poster Session, (page 16)]

Yuting CHEN, *University of Maryland, College Park*

There is a growing demand to improve on direct estimates of parameters for small geographical areas or domains where little or no sample is available from the primary data source. Estimation methods that use linear mixed models to combine information from multiple data sources have been used to address such small area estimation problems. In this paper, using Monte Carlo simulation experiments, we study the effect of model misspecifications on a few commonly used small area point estimators and their associated mean squared prediction error estimators.

62. Factor Analysis of Data with Incomplete Records

[03.A1.I39, (page 25)]

Fan DAI, *Michigan Technological University*
Somak DUTTA, *Iowa State University*
Ranjan MAITRA, *Iowa State University*

Data with partial records arise naturally in modern applications as a result of the collection and preprocessing process, usually causing difficulties in the implementation of standard techniques for analyzing the data variability. We develop exploratory factor analysis for incomplete data to characterize the vari-

ability in features in terms of a few interpretable latent factors. The maximum likelihood estimates of model parameters are provided by a novel conditional expectation-maximization algorithm with matrix-free computations. Results from the simulation studies show the superiority of our algorithm. Our method applied to gamma-ray burst data, capturing records of hand postures, and semiconductor functionality testing results provides insights into the underlying mechanisms that determine the variability in each dataset.

63. Using continuous methane measurements for inventory development on oil and gas sites: three case studies

[Poster Session, (page 16)]

William, S DANIELS, *Colorado School of Mines*

Creating accurate methane emissions tallies ("inventories") for the oil and gas sector is a critical aspect of recent environmental regulation in the United States (e.g., the Inflation Reduction Act). The current method for creating inventories utilizes empirically-derived, static emission factors that have proven to be inadequate and severely underestimate methane emissions in general. To improve these inventories, much research has gone into developing measurement-informed inventories, often using snapshot, top-down methane measurements from a plane or drone. One inherent limitation of these top-down approaches is that methane emissions have a high degree of temporal variability, which makes it challenging to produce accurate inventories using only a small number of snapshot measurements. Here we present three case studies that demonstrate how methane emission estimates derived from ground-based continuous methane monitoring sensors can be used jointly with snapshot measurements to create accurate, measurement-informed emissions inventories.

64. Variable selection in quantile regression with ordinal responses

[04.M2.I57, (page 33)]

Mai DAO, *Wichita State University*
Md Sakhawat HOSSAIN, *Texas Tech University*

Since the pioneering work of Koenker and Bassett (1978), quantile regression has been a popular regression technique that helps researchers investigate the whole distribution of the response variable. In practice, ordinal responses appear frequently and have significant importance in many applications. In this talk, we discuss a Bayesian hierarchical model to conduct parameter estimation and variable selection for quantile regression with such outcomes. We use the latent response variable generated by the ordinal outputs and the mixture representation of the asymmetric Laplace distribution to set up the quantile regression model. Then, we employ the horseshoe prior and group sampling of the latent dependent variable and cutpoints to reduce autocorrelation in the generated posterior samples in high-dimensional settings. Finally, we utilize the sequential two-means clustering process to select important predictors for the model selection and comparison. We conduct both simulation studies and real data applications to illustrate the feasibility and computational advantages of the proposed algorithm.

65. Multi-scale Genome-wide Mediation Analysis (M-GMAS) in Twin Imaging Studies

[Poster Session, (page 16)]

Anisha DAS, *Florida State University*

This study develops a computationally efficient model for twin data that can incorporate multiple phenotypes at a single run. The data comprises of information on monozygotic and dizygotic twins as well as singletons, which has been split into two groups and the ACE model has been executed to obtain the dynamic heritabilities and decide on the degree of association between a given SNP and phenotypic responses. Then the Twin GWAS analysis is performed that gives us the significant SNPs and their relation with the phenotypes in hand. This is a significant effort to use the ACEt and the TwinEQTL packages in R software on phenotypic response data in matrix format. Techincally, this is the first time that Twin GWAS analysis is applied on multiple phenotypes at one go. The study also determines the polygenic risk scores for the significant SNPs obtained from Twin GWAS analysis, and use these in mediation analysis for figuring out the causal relationship between certain chosen clinical outcomes and phenotypes.

Keywords: Twin GWAS, SNPs, ACE model, mediation analysis

66. Clustering sequence data with mixture Markov chains with covariates using multiple simplex constrained optimization routine

[04.M1.I55, (page 32)]

Priyam DAS, *Virginia Commonwealth University*
Deborshee SEN , *Amazon, India*
Debsurya DE, *Johns Hopkins University*
Tianxi CAI, *Harvard Medical School*

Mixture Markov Model (MMM) is a widely used tool to cluster sequences of events coming from a finite state-space. However, the MMM likelihood being multi-modal, the challenge remains in its maximization. Although Expectation-Maximization (EM) algorithm remains one of the most popular ways to estimate the MMM parameters, however convergence of EM algorithm is not always guaranteed. Given the computational challenges in maximizing the mixture likelihood on the constrained parameter space, we develop a pattern search-based global optimization technique which can optimize any objective function on a collection of simplexes, which is eventually used to maximize MMM likelihood. This is shown to outperform other related global optimization techniques. In simulation experiments, the proposed method is shown to outperform the expectation-maximization (EM) algorithm in the context of MMM estimation performance. The proposed method is applied to cluster Multiple sclerosis (MS) patients based on their treatment sequences of disease-modifying therapies (DMTs). We also propose a novel method to cluster people with MS based on DMT prescriptions and associated clinical features (covariates) using MMM with covariates. Based on the analysis, we divided MS patients into 3 clusters. Further cluster-specific summaries of relevant covariates indicate patient differences among the clusters.

67. Blocked Gibbs sampler for hierarchical Dirichlet processes

[Poster Session, (page 16)]

Snigdha DAS, *Department of Statistics, Texas A&M University*

Posterior computation in hierarchical Dirichlet process (HDP) mixture models is an active area of research in nonparametric Bayes inference of grouped data. Existing literature almost exclusively focuses on the Chinese restaurant franchise (CRF) analogy of the marginal distribution of the parameters, which can mix poorly and is known to have a linear complexity with the sample size. A recently developed slice sampler allows for efficient blocked updates of the parameters, but is shown to be statistically unstable in our article. We develop a blocked Gibbs sampler to sample from the posterior distribution of HDP, which produces statistically stable results, is highly scalable with respect to sample size, and is shown to have good mixing. The heart of the construction is to endow the shared concentration parameter with an appropriately chosen gamma prior that allows us to break the dependence of the shared mixing proportions and permits independent updates of certain log-concave random variables in a block. En route, we develop an efficient rejection sampler for these random variables leveraging piece-wise tangent-line approximations.

68. Hierarchical Bayes estimation of small area proportions using statistical linkage of disparate data sources

[Poster Session, (page 16)]

Soumojit DAS, *University of Maryland, College Park*

We propose a Bayesian approach to estimate finite population proportions for small areas. The proposed methodology improves on the traditional sample survey methods because, unlike the traditional methods, our proposed method borrows strength from multiple data sources. Our approach is fundamentally different from the existing small area Bayesian approach to the finite population sampling, which typically assumes a hierarchical model for all units of the finite population. We assume such model only for the units of the finite population in which the outcome variable observed; because for these units, the assumed model can be checked using existing statistical tools. Modeling unobserved units of the finite population is chal- lenging because the assumed model cannot be checked in the absence of data on the outcome variable. To make reasonable modeling assumptions, we propose to form a large number of cells for each small area using factors that potentially influence the binary outcome variable of interest. This strategy is expected to bring some degree of homogeneity within a given cell and also among cells from different small areas that are constructed with the same factor level combination. Instead of modeling true probabilities for unobserved individual units, we assume that population means of cells with the same combination of factor levels are identical across small areas and the population mean of true probabilities for a cell is identical to the mean of true values for the observed units in that cell. We apply our proposed methodology to a real-life COVID-19 survey, linking information from multiple disparate data sources to estimate vaccine-hesitancy rates (proportions) for 50 US

states and Washington, D.C. (small areas). We also provide practical ways of model selection that can be applied to a wider class of models under similar setting but for a diverse range of scientific problems.

69. The power of r-power

[01.E1.I24, (page 14)]

Nairanjana DASGUPTA, *Washington State University*

Large scale multiplicity is an oft-encountered problem in medical, pharmaceutical, ecological, and engineering studies. The idea is how to balance multiple tests with the probability of false positives. In the recent years False Discovery Rates, FDR, has captured a lot of interest among practitioners. As an alternative Dasgupta et. al. (2016) introduced a concept r-power is to determine the size of lists or number of "positives" a practitioner would have resources to follow-up on and then look at the reliability of such a list. In this talk we will look at the idea of r-power and talk about it in terms of the issues around the normal means problem as it relates to r-power.

70 . A Pseudo-Bayesian Approach to Small Area Estimation Using Spatial Models

[03.A2.I44, (page 27)]

Gauri DATTA, *Univ of Georgia and US Census Bureau*
Jiacheng LI, *Univ of Georgia*

The empirical best linear unbiased prediction (EBLUP) method has been the dominant frequentist model based approach in small area estimation. As an alternative to the EBLUP method in small area estimation, the observed best prediction (OBP) method was proposed by Jiang et al. (2011) where the parameters of the small area model are estimated by minimizing an objective function that is implied by the total mean squared prediction error. In a recent article, Datta, Lee and Li (2023) followed a general Bayesian approach developed by Bissiri et al. (2016) to develop a pseudo-Bayesian method by appropriately calibrating the objective function of the OBP method for the Fay-Herriot model. In the absence of suitable covariates with good predictive power the small area estimates from the standard Fay-Herriot model can be improved by using spatially dependent random effects. In this talk, we present a pseudo-Bayes SAE method using spatial Fay-Herriot model. Evaluation of the proposed method based on an application to estimation of four-person family median incomes for the U.S. states shows the usefulness of the proposed method.

71. Data Adaptive Covariate Balancing for Causal Effect Estimation for High Dimensional Data

[03.M2.C1, (page 22)]

Simion DE, *Biostatistics PhD Student, University of Minnesota*
Jared HULING, *Assistant Professor,University of Minnesota*

A critical task in accurately estimating causal effects from observational data is to account for confounding, often achieved by weighting techniques aimed at balancing the distribution of confounders between the treated and control groups. Weighting techniques can be classified in two ways. The first way is based on whether estimating weight is parametric or non-parametric. The second way is based on whether one 1) models the propensity score and inverts it or 2) directly constructs weights that attempt to achieve distributional balance between the treated and the control groups. Parametric methods, both modeling and direct balancing, suffer from model misspecification while balancing techniques suffer from the curse of dimensionality. Methods have been developed to break the curse of dimensionality by identifying confounders among many candidate variables. But these are parametric and focus on modeling the propensity score and therefore subject to bias. In this paper, we propose a nonparametric direct balancing approach that uses a random forest to data-adaptively balance on confounders. Our method uses random forests to jointly model the outcome and treatment based on the covariates. To construct a measure of distributional balance that emphasizes covariates that impact both treatment and outcome, we propose a distance based on the proportion of trees in which two observations appear in the same leaf node, resulting in a distance, sensitive to confounders that can reduce dimensionality while focusing directly on the source of bias in estimating a causal effect. We demonstrate the highly competitive performance of our method using extensive simulations.

72. Estimating heterogeneous treatment effects on binary outcomes with non-compliance using Bayesian additive re-

gression trees

[01.M1.I5, (page 5)]

Sameer DESHPANDE, *University of Wisconsin–Madison*
Jared FISHER, *Brigham Young University*
David PUELZ, *University of Texas at Austin*

Estimating varying treatment effects in randomized trials with noncompliance is inherently challenging since variation comes from two separate sources: variation in the impact itself and variation in the compliance rate. In this setting, existing Frequentist and ML-based methods are quite flexible but are highly sensitive to the so-called weak instruments problem, in which the compliance rate is (locally) close to zero, and require pre-specifying subgroups of interest. Parametric Bayesian approaches, which account for noncompliance via imputation, are more robust in this case, but are much more sensitive to model specification. In this paper, we propose a Bayesian semiparametric approach that combines the best features of both approaches. Our main methodological contribution is to present a Bayesian Causal Forest model for binary response variables in scenarios with noncompliance. In this Bayesian noncompliance framework, we repeatedly impute individuals' compliance types, allowing us to flexibly estimate varying treatment effects among compliers while mitigating the weak instruments problem. We then apply the method to detect and analyze heterogeneity in study of workplace wellness, where there are a plethora of binary outcomes of interest.

73. Robustifying likelihoods by optimistically re-weighting data

[01.A2.I20, (page 12)]

Miheer DEWASKAR, *Duke University*
Christopher TOSH, *Memorial Sloan Kettering Cancer Center*
Jeremias KNOBLAUCH, *UCL*
David DUNSON, *Duke University*

Likelihood-based inferences have been remarkably successful in wide-spanning application areas. However, even after due diligence in selecting a good model for the data at hand, there is inevitably some amount of model misspecification: outliers, data contamination or inappropriate parametric assumptions such as Gaussianity mean that most models are at best rough approximations of reality. A significant practical concern is that for certain inferences, even small amounts of model misspecification may have a substantial impact; a problem we refer to as brittleness. In this talk, we address the brittleness problem in likelihood-based inferences by choosing the most model friendly data generating process in a discrepancy-based neighbourhood of the empirical measure. This leads to a new Optimistically Weighted Likelihood (OWL), which robustifies the original likelihood by formally accounting for a small amount of model misspecification. Focusing on total variation (TV) neighborhoods, we study theoretical properties, develop inference algorithms and illustrate the methodology in applications to mixture models and regression.

Work based on manuscript:
https://arxiv.org/abs/2303.10525

74. Generalized Variable Selection Algorithms for Gaussian Process Models by LASSO-like Penalty

[04.M1.I51, (page 31)]

Dipak DEY, *University of Connecticut*
Zhiyong HU, *Microsoft New England Research and Development Center*

With the rapid development of modern technology, massive amounts of data with complex patterns are generated. Gaussian process models that can easily fit the nonlinearity in data become more and more popular nowadays. It is often the case that in some data only a few features are important or active. However, unlike classical linear models, it is challenging to identify active variables in Gaussian process models. One of the most used methods for variable selection in Gaussian process models are automatic relevance determination, which is known to be open-ended. There is no rule of thumb to determine the threshold for dropping features, which makes the variable selection in Gaussian process models ambiguous. In this work, we propose two variable selection algorithms for Gaussian process models, which use the artificial nuisance columns as baseline for identifying the active features. Moreover, the proposed methods work for both regression and classification problems. The algorithms are demonstrated using comprehensive simulation experiments and an application to multi-subject electroencephalography (EEG) data that studies alcoholic levels of experimental subjects.

Keywords: Automatic relevance determination, Electroencephalography data, Gaussian process, Principal component analysis, Variable selection

75 . Assessing Contribution of Treatment Phases through Tipping Point Analyses via Counterfactual Elicitation Using Rank Preserving Structural Failure Time Models

[03.A2.I42, (page 26)]

Jyotirmoy DEY, *Regeneron Pharmaceuticals, Inc.*
Sudipta BHATTACHARYA, *Takeda Pharmaceuticals, Inc.*

This article provides a novel approach to assess the importance of specific treatment phases within a treatment regimen through tipping point analyses (TPA) of a time-to-event endpoint using rank-preserving structural-failure-time (RPSFT) modeling. In oncology clinical research, an experimental treatment is often added to the standard of care therapy in multiple treatment phases to improve patient outcomes. When the resulting new regimen provides a meaningful benefit over standard of care, gaining insights into the contribution of each treatment phase becomes important to properly guide clinical practice. New statistical approaches are needed since traditional methods are inadequate in answering such questions. RPSFT modeling is an approach for causal inference, typically used to adjust for treatment switching in randomized clinical trials with time-to-event endpoints. A tipping-point analysis is commonly used in situations where a statistically significant treatment effect is suspected to be an artifact of missing or unobserved data rather than a real treatment difference. The methodology proposed in this article is an amalgamation of these two ideas to investigate the contribution of a specific component of a regimen comprising multiple treatment phases. We provide different variants of the method and construct indices of contribution of a treatment phase to the overall benefit of a regimen that facilitates interpretation of results. The proposed approaches are illustrated with findings from a recently concluded, real-life phase 3 cancer clinical trial. We conclude with several considerations and recommendations for practical implementation of this new methodology.

76 . Policy evaluation in reinforcement learning: The impact of temporal dependence and multi-step lookahead

[04.M2.I61, (page 34)]

Yaqi DUAN, *MIT*
Martin WAINWRIGHT, *MIT*

In this talk, we investigate non-parametric estimation of the value function for a Markov reward process. Our first focus is on the impact of temporal dependence in the data. Our theory reveals that there are delicate interactions between mixing and model mis-specification. Surprisingly, when the model is well-specified, the correlation of the data does not reduce the quality of estimation. In contrast, when the model is significantly mis-specified, temporal dependence may deteriorate the performance of the temporal difference estimate. In the second part of the talk, we study the choice of lookahead in multi-step temporal difference learning. Our theory suggests that the bias-variance tradeoff, which has been long discussed in classical textbooks, may not necessarily occur. Specifically, the variance is not sensitive to the changes in lookahead. Moreover, if the model is mis-specified and data mixing is significant, increasing lookahead can lead to a decrease the bias and an improvement in the quality of the estimate.

77 . Using External Control Arm to Benchmark Time-to-Event Outcomes in Single-Arm Trials: A Case Study on Triple Negative Breast Cancer Patients

[01.E1.I23, (page 13)]

Abhishek DUBEY, *Bristol Myers Squibb*
Arun KUMAR, *Bristol Myers Squibb*
Kaushal MISHRA, *Bristol Myers Squibb*
Kalyanee VIRASWAMI-APPANNA, *Bristol Myers Squibb*
Armand CHOUZY, *Bristol Myers Squibb*
Ram TIWARI, *Bristol Myers Squibb*

Many Phase 1/2 clinical trial lack a concurrent control arm, making it challenging to interpret study results. To address this issue, we have developed a synthetic control arm (SCA) using standard of Care data available from real world database (Flatiron Health†) to interpret a cohort of triple negative breast cancer (TNBC) patients from a single-arm trial. TNBC is an aggressive form of breast cancer and has an unmet medical need due to poor prognosis; 5-year survival rate of 12% for patients with advanced disease [Ref. 1]. A SCA using an outcome-free propensity score-based algorithm was created from 407 TNBC patients with locally advanced or metastatic disease with ECOG score < 2 in the Flatiron Health real-world data (RWD). The algorithm matched important covariates, such as ECOG score,

immunotherapy history, site of metastasis, and others from the experimental arm with those in the RWD. The median progression-free survival (PFS) and overall survival with 95% confidence intervals (CI) were estimated using Kaplan Meier method and bootstrapping for the SCA. The outcome-free matching algorithm resulted in similar baseline characteristics in SCA and experimental arm; a low Maximum Mean Discrepancy of 0.06 is achieved. The median PFS was estimated to be 3.10 months (95% CI 2.83, 4.08), which will be used as benchmark to assess PFS data for experimental arm when the data becomes available. Creating the SCA presented challenges, such as identifying the covariates needed to be balanced between the SCA and experimental arm, balancing a large number of covariates, determining the number of patients to be selected in the SCA or determining their weights when all patients included, choosing an appropriate data cut strategy to ensure comparability of follow-up duration, and disparity in PFS definition used in trial and RWD. Our work highlights how RWD data can be used to benchmark time-to-event outcomes from single-arm trials despite the challenges and can inform decision-making in Phase 1/2 proof-of-concept oncology studies.

78. Spectral Universality in Regularized Linear Regression with Nearly Deterministic Designs

[03.E1.I46, (page 28)]

Rishabh DUDEJA, *Harvard University*
Subhabrata SEN, *Harvard University*
Yue M. LU, *Harvard University*

Spectral universality refers to the empirical observation that asymptotic properties of a high-dimensional stochastic system driven by a structured random matrix are often determined only by the spectrum (or singular values) of the underlying matrix - the singular vectors are irrelevant provided they are sufficiently "generic". Consequently, the properties of the underlying system can be accurately predicted by analyzing the system under the mathematically convenient assumption that the singular vectors as uniformly random (or Haar distributed) orthogonal matrices. This general phenomenon has been observed in numerous contexts, including statistical physics, communication systems, signal processing, statistics, and randomized numerical linear algebra. We study this universality phenomenon in the context of high-dimensional linear regression, where the goal is to estimate an unknown signal vector from noisy linear measurements specified using a design matrix. We prove a spectral universality principle for the performance of convex regularized least squares (RLS) estimators for this problem. Our contributions are two-fold: (1) We introduce a notion of a universality class for design matrices, defined through nearly deterministic conditions that fix the spectrum of the design and formalize the heuristic notion of generic singular vectors; (2) We show that for all design matrices in the same universality class, the dynamics of the proximal gradient algorithm for the regression problem, and the performance of RLS estimators themselves (under additional strong convexity conditions) are asymptotically identical. In addition to including i.i.d. Gaussian and rotational invariant matrices as special cases, our universality class also contains highly structured, strongly dependent, and even nearly deterministic designs. Examples include randomly signed incoherent tight frames and randomly subsampled Hadamard transforms. Due to this universality result, the performance of RLS estimators on many structured design matrices with limited randomness can be characterized using the rotationally invariant design with uniformly random (or Haar distributed) singular vectors as an equivalent yet mathematically tractable surrogate.

79. Sparse canonical correlation for integrative multi-omics explain pan-cancer and cancer-specific patterns of association

[01.A1.I13, (page 9)]

Diptavo DUTTA, *NIH/NCI*

Complex diseases like cancer involve intricate downstream regulation of molecular phenotypes like gene expression, protein levels by disease-related variants. However, such molecular targets have been shown to have noticeable overlap across types of cancers as well as demonstrate cancer specific patterns. Here we introduce a canonical correlation-based method for low-rank decomposition of associations of cancer-related genetic variants with molecular phenotypes. Starting with publicly available summary statistics on the association of cancer-related genetic variants with thousands of molecular targets like gene expressions or protein levels, we use low-rank approximations to identify shared as well as cancer-specific sets of molecular targets. Further, using a competitive hypothesis, we test the enrichment of the identified target sets against a general polygenic background. Analysis of available eQTLGen

data across 8 cancer types, replicate several known genes and reveal novel targets. Further, using genetic variants associated with 21 different cancers and plasma protein levels, we identify shared proteins across known groups of cancers as well as cancer specific protein targets.

80. Matrix-free maximum likelihood estimation of Gaussian factor models

[03.A1.I39, (page 25)]

Somak DUTTA, *Iowa State University*
Ranjan MAITRA, *Iowa State University*
Fan DAI, *Michigan Technological University*

Abstract: Factor models are often useful in characterizing the dependence among several variables using a few latent factors. The maximum likelihood estimators for the parameters of a Gaussian factor model are estimated through the profile likelihood function. However, this profile likelihood function does not work when the number of variables is larger than the sample size. In this talk, we extend the profile likelihood method to high-dimensional Gaussian data in terms of partial singular value decomposition. By implementing a restarted Lanczos algorithm and a limited-memory quasi-Newton method, we develop a matrix-free algorithm for fast computations and gain substantial speedup over the EM algorithm. We illustrate the method on a high-dimensional fMRI dataset from a study on suicide attempters, suicide ideators, and a control group.

81. How to learn more about practical aspects of clinical trials

[01.E1.I22, (page 13)]

Dixie ECKLUND, *University of Iowa*

TBA

82. Statistical data integration using multilevel models

[01.A2.I18, (page 11)]

Andreea ERCIULESCU, *Westat*
Jean OPSOMER, *Westat*
Benjamin SCHNEIDER, *Westat*

This article considers the case where two surveys collect data on a common variable, with one survey being much smaller than the other. The smaller survey collects data on an additional variable of interest, related to the common variable collected in the two surveys, and out-of-scope with respect to the larger survey. Estimation of the two related variables is of interest at domains defined at a granular level. We propose a multilevel model for integrating data from the two surveys, by reconciling survey estimates available for the common variable, accounting for the relationship between the two variables, and expanding estimation for the other variable, for all the domains of interest. The model is specified as a hierarchical Bayes model for domain-level survey data, and posterior distributions are constructed for the two variables of interest. A synthetic estimation approach is considered as an alternative to the hierarchical modelling approach. The methodology is applied to wage and benefits estimation using data from the National Compensation Survey and the Occupational Employment Statistics Survey, available from the Bureau of Labor Statistics, Department of Labor, United States.

83. ECoHeN: A Hypothesis Testing Framework for Extracting Communities from Heterogeneous Networks

[03.M1.I30, (page 20)]

Bailey FOSDICK, *Colorado School of Public Health*
Connor GIBBS, *Colorado State University*
James WILSON, *University of San Francisco*

Community discovery is the general process of attaining assortative communities from a network: collections of nodes that are densely connected within the collection yet sparsely connected to the rest of the network. While community discovery is well developed in a simple network setting, few such techniques exist for heterogeneous networks, which contain different types of nodes and possibly different connectivity patterns between the node types. As many complex phenomena can be represented as a heterogeneous network, great interest lies in uncovering communities that are topologically dense with respect to the node type of its' members. In this talk, we introduce a framework called ECoHeN, which extracts communities from a heterogeneous network in a statistically meaningful way. Using a heterogeneous configuration model as a reference distribution, ECoHeN identifies communities that are more densely connected than expected given the node types and connectivity of its membership. Specifically, ECoHeN extracts communities one at a time through a dynamic set of iterative updating rules, which are guaranteed to converge, and where the extracted community may contain nodes with similar or dissimilar node type. To our knowledge this is the first discovery method that distinguishes and uncovers both ho-

mogeneous and heterogeneous, possibly overlapping, community structure in a network.

84. Combining multiple sources of information to estimate hearing loss prevalence in the United States at the county level by gender, age, and race/ethnicity using small area estimation models

[01.M1.I5, (page 4)]

Carolina FRANCO, *NORC at the University of Chicago*
David REIN,

We discuss the statistical methodology for the ambitious undertaking of producing estimates of hearing loss for US counties for six age groups, two genders, and four race/ethnicity groups. The new estimates can inform interventions and planning at the county level. Our small area estimation modeling strategy uses data from the National Health and Nutrition Examination Survey (NHANES) and the American Community Survey (ACS), and auxiliary information from Medicare claims diagnoses, The Area Health Resources File, and industry data on audiologists. NHANES provides hearing loss estimates measured by trained technicians that capture different severities of hearing loss. In contrast, ACS provides binary and self-reported estimates of any hearing loss, which are less accurate than the NHANES instrument. However, NHANES' sample size of about 5000 persons per year is too small for county estimates, even when combining multiple years of data, while ACS samples about 3.5 million addresses yearly. The methodology we introduce uses small area estimation models to capture hearing loss as measured by the two surveys, using covariates from auxiliary data to borrow strength. We use calibration to national NHANES modeled totals by demographic groups to correct the self-report bias of county modeled ACS hearing loss estimates. Using this approach, implemented from a Bayesian paradigm via MCMC, we estimate two severity levels of hearing loss at the county and demographic levels. Our methodology incorporates the sampling error from all survey sources when computing uncertainty measures. We presents strategy to combine information from multiple sources and calculate uncertainty. The model results provide previously unavailable information about hearing loss prevalence at the county level, which illustrates differences by race and ethnicity and levels of urbanization.

85 . Sketched Gaussian Processes: A Strategy for Predictive Inference for High-Dimensional Features

[Poster Session, (page 16)]

Samuel F. GAILLIOT, *Texas A&M University*

We focus on drawing computationally-efficient predictive inference from Gaussian process (GP) regressions with a large number of features when the response is conditionally independent of the features given the projection to a noisy low dimensional manifold. Bayesian estimation of the regression relationship using Markov Chain Monte Carlo and subsequent predictive inference is computationally prohibitive and may lead to inferential inaccuracies since accurate variable selection is essentially impossible in such high-dimensional GP regressions. As an alternative, this article proposes a strategy to sketch high-dimensional feature vectors, before fitting a GP with the scalar outcome and the sketched feature vector. The analysis is performed in parallel with many different random sketching matrices and smoothing parameters, and the predictive inferences are combined using Bayesian predictive stacking. The algorithm allows fast implementation with very large p and offers theoretical support on the accuracy of prediction. Empirical studies show superior performance of the proposed approach with a wide variety of competitors.

86 . Likelihood-based spatiotemporal forecasting of burned area due to wildfire

[Poster Session, (page 16)]

Indrila GANGULY, *North Carolina State University*

In the last few decades, wildfires have increased in frequency. Such increased frequencies can be attributed to climate change, which has been a major contributor to several natural disasters in the recent past. Increased wildfires potentially lead to significant health hazards for humans. Hence, a relevant problem today is to be able to forecast the burned area from wildfires, which can be subsequently used to forecast the effects of wildfire on human health. However, we do not find many works in the literature focusing on wildfire prediction in real-time, and predicting the total burned area. In our work, we have developed a new approach for modeling wildfire burned area, combining the ideas of cellular automata models and models on infectious disease spread. The parameters of our model take into account the effects

of covariates, including weather conditions and wind speed that can affect the spread and stopping of the fire. The model is trained on observed wildfire data to estimate parameters, which can then be used for forecasting future burned area. We demonstrate the empirical performance of our model via a simulation study and show its applications to California wildfire data.

87. A Bayesian approach for utilizing historical data in early drug development

[01.A2.I20, (page 12)]

Nairita GHOSAL, *Merck & Co., Inc., Rahway, NJ, USA*

Humankind has made tremendous progress in developing therapeutics and vaccines in the last hundred years. This leads to the reality that a newly developed treatment often has a predecessor and might have slight incremental benefits over existing treatments. In noninferiority trial settings, large amounts of clinical data are often accessible for the study's control arm. In early drug development where sample sizes are generally small, historical data can provide substantial information on a control arm in order to provide more resources towards the novel treatment. Historical borrowing can also provide better point estimates and increase power in clinical trials. This presentation investigates relative risk of adverse events between the treatment and the control arm, considering historical borrowing in the control arm. A brief description of using historical borrowing in early drug development will be provided using simulated examples.

88. Limit theorems for high dimensional least-square online SGD

[04.M2.I61, (page 35)]

Promit GHOSAL, *Massachusetts Institute of Technology*
Bhavya AGRAWALLA, *Massachusetts Institute of Technology*
Krishnakumar BALASUBRAMANIAN, *University of California, Davis*
Ye HE, *University of California, Davis*

Stochastic gradient descent (SGD) has emerged as the quintessential method in a data scientist's toolbox. Much progress has been made in the last two decades toward understanding the iteration complexity of SGD (in expectation and high-probability) in the learning theory and optimization literature. However, using SGD for high-stakes applications requires careful quantification of the associated uncertainty. Toward that end, in this talk, we discuss high-dimensional Central Limit Theorems (CLTs) for linear functionals of online least-squares SGD iterates under a Gaussian design assumption. Furthermore, we describe the fluctuation and scaling limit of the whole trajectory of high dimensional online least-squares SGD.

89. The envelope distribution of a complex Gaussian random variable

[Poster Session, (page 16)]

SATTWIK GHOSAL, *IOWA STATE UNIVERSITY*

This article explicitly derives the cumulative distribution function of the envelope of a elliptical Gaussian complex vector, or equivalently, the norm of a bi variate normal random vector with general covariance structure. Some properties of the distribution, specifically, its moments and moment generating functions, are also derived and shown to exist. These functions and expressions are exploited to reduce to those for the special case distributions.

90. A Bootstrap based Goodness-of-fit test of covariance for multiple outcomes in Longitudinal Data

[01.A1.I12, (page 8)]

Dhrubajyoti GHOSH, *Duke University*
Sheng LUO, *Duke University*

The practice of using functional data methods has long been frequent in longitudinal data analysis. While they offer significant flexibility in terms of capturing dependence across repeated observations, it is computationally intensive and is much more complex compared to using the parametric covariance function. We have proposed a multivariate goodness of fit test, based on the previous works by Chen et al. (2019) for univariate data, utilizing max-based test statistics and used bootstrap to provide an estimate of the distribution of test statistic under the null hypothesis. Our proposed method contributes a way to test whether a linear mixed-effects model is sufficient for a given data, or whether a more complicated set-up is required, by testing a quadratic polynomial structure induced by a linear mixed-effects model. We have provided simulation studies using both regular and irregularly spaced data points, and two real data analyses from ADNI and MDS-UPDRS to illus-

trate the test's performance.

91. Online Bayesian Variable Selection for Streaming Data

[01.E1.I21, (page 13)]

Joyee GHOSH, *The University of Iowa*
Aixin TAN, *The University of Iowa*

There are many modern applications in which data are continuously generated and one would like to learn from the data sequentially, after each dataset arrives. For example, such streaming datasets can be generated by wearable devices such as fitbit or apple watch, or devices for monitoring pollution or health. The main challenge with such datasets is they are too large for computer memory and traditional statistical methods can be extremely slow or impossible to implement. Renewable estimation has been shown to have excellent properties for online estimation in frequentist methods. This method stores some key summary statistics from previous datasets, instead of storing all individual records or in other words the entire stream of data. In this work, we exploit the idea of renewable estimation for developing algorithms for Bayesian variable selection for generalized linear models for streaming datasets.

92 . Interrelationship between Divergence Measures

[Memorial Session 1, (page 12)]

Malay GHOSH, *University of Florida*
Partha SARKAR, *University Of Florida*

The paper considers a number of measures typically used for divergence between two distributions, and studies their interrelationship.

93. Variable selection in non-linear metric learning

[04.M2.I57, (page 33)]

Souparno GHOSH, *University of Nebraska-Lincoln*

In complex setting, metric learning offers a way to estimate the distance metric which can be subsequently fed into standard K-nearest neighbor type prediction methods. Standard metric learning methods do not consider the variable selection problem because reducing the dimension of the coordinate system in a pointwise fashion may not be meaningful. However, when metric learning is deployed in a setting with spurious predictors, the performance of the subsequent prediction mechanism suffers. We develop a sequential algorithm to perform variable selection in the context of non-linear ordinal metric learning and demonstrate its efficacy in cancer drug response data.

94. A robust Kernel Machine framework for assessing differential expression of multi sampled single cell data

[03.M1.I31, (page 20)]

Tusharkanti GHOSH, *Colorado School of Public Health*
Debashis GHOSH, *Colorado School of Public Health*

n this talk, we present cytoKernel, a robust method for differential expression of single-cell data using a kernel-based score test. It is specifically designed to assess the differential expression of single cell RNA sequencing and high-dimensional flow or mass cytometry data using the full distributions. High-throughput sequencing of single-cell data provides a rigorous view of cell specification, enabling intricate variations between groups or conditions to be identified. However, many existing methods for differential expression only target differences in aggregate measurements and limit their approaches to detect only global differential changes. In contrast, cytoKernel is based on a semi-parametric logistic regression model that employs the full distributions of the single cell data. By calculating the divergence between pairwise distributions of subjects, it can detect both differential patterns involving changes in the aggregate, as well as more elusive variations that are often overlooked due to the multimodal characteristics of several single cell data. We performed extensive benchmarks across both simulated and real data sets from single cell mass cytometry data and RNA sequencing. The results show that cytoKernel effectively controls the False Discovery Rate (FDR) and identifies more differential patterns than existing approaches. We also applied cytoKernel to assess gene expression and protein marker expression differences from cell subpopulations in various single cell RNAseq and mass cytometry data sets.

95 . Employing Tensor Regression for Analyzing the Effect of Alcohol on the Brain

[Poster Session, (page 16)]

Clarissa L. GIEFER, *South Dakota State University*

Excessive alcohol consumption is one of the leading causes of premature deaths in the United States.

Neuroimaging techniques can be used to locate the areas of an alcoholics' brain that are affected by drinking. The purpose of this research was to compare different tensor decomposition methods and studied their quality of fit and prediction for analysis. This was done by comparing diagnosed alcoholic brain imaging results against those of healthy controls using diffusion tensor imaging (DTI) from a group of 89 alcoholics and 89 non-alcoholic healthy controls (age 22–36, 88 female) from Human Connectome Project (HCP). This discussion focuses on two of the main tensor decompositions, Tucker decomposition, and canonical decomposition/parallel factors (CANDECOMP/PARAFACE), better known as CP decomposition. Tensor response regression was used to detect areas of the brain that are activated by the predictors. The effectiveness of CP regression and Tucker partial least squares (PLS) and 1D regression were compared to the ordinary least squares method to demonstrate their predictive accuracy. Compared to OLS, the Tucker PLS and 1D methods had similar results performing better than OLS. A significant reduction in error was reported under the CP regression with more commonly occurring activation areas found.

96. Bayesian model-based synthetic control methods

[03.M2.I32, (page 22)]

Gyuhyeong GOH, *Kansas State University*
Jisang YU, *Kansas State University*

Synthetic Control Method (SCM) for casual inference with a single treated unit has received great attention in recent years. The SCM framework is commonly known as a data-driven method since it constructs the so-called synthetic control unit based on the convex hull of the untreated units. Although several inference procedures for SCM have been proposed, there is a strong need for methods of formal statistical inference. In this study, we develop a model-based SCM approach that allows us to implement Bayesian statistical inference. From our model-based SCM point of view, we show that the conventional SCM can be considered as a special case when the model is overfitted. To address the problem of overfitting, we propose to exploit global-local shrinkage priors that induce rank reduction and sparsity in the synthetic control unit. The performance of the proposed Bayesian SCM is investigated via extensive simulation studies.

97. Multivariate Spatial Prediction of Air Pollutants

[03.A1.I37, (page 24)]

Wenlong GONG, *University of Houston - Downtown*
Brian REICH, *North Carolina State University*
Howard CHANG, *Emory University*

Estimates of daily air pollution concentrations with complete spatial and temporal coverage are important for supporting epidemiologic studies and health impact assessments. While numerous approaches have been developed for modeling air pollution, they typically only consider each pollutant separately. We describe a spatial multipollutant data fusion model that combines monitoring measurements and chemical transport model simulations that leverages dependence between pollutants to improve spatial prediction. For the contiguous United States, we created a data product of daily concentration for 12 pollutants (CO, NOx, NO2, SO2, O3, PM10, and PM2.5 species EC, OC, NO3, NH4, SO4) during the period 2005 to 2014. Out-of-sample prediction showed good performance, particularly for daily PM2.5 species EC ($R2 = 0.64$), OC ($R2 = 0.75$), NH4 ($R2 = 0.84$), NO3 ($R2 = 0.73$), and SO4 ($R2 = 0.80$). By employing the integrated nested Laplace approximation (INLA) for Bayesian inference, our approach also provides model-based prediction error estimates. The daily data product at 12 km spatial resolution will be publicly available immediately upon publication. To our knowledge this is the first publicly available data product for major PM2.5 species and several gases at this spatial and temporal resolution.

98. Post-selection inference for regression with grouped responses

[01.A1.I12, (page 8)]

Karl GREGORY, *University of South Carolina*
Qinyan SHEN, *University of South Carolina*
Xianzheng HUANG, *University of South Carolina*

We consider adjusting inferences to account for model selection in regression with grouped responses, in particular in logistic regression when the individual response values (disease statuses) are not observed; instead, information about the responses comes from collections of error prone tests on groups of individuals. We extend recently developed post-LASSO-selection inference methods based on the "polyhedral lemma." We find that the polyhedral lemma method works well in our setting with grouped responses, pro-

vided the level of sparsity penalization is chosen independently of the data. This caveat makes the implementation of post-LASSO-selection inference methods based on the polyhedral lemma problematic, as they assume the level of sparsity penalization to be given; in practice, one must "dip" twice into the data to use these methods — once to select the level of penalization and once more to perform estimation and inference.

99. Scalable Nonparametric Bayesian Learning for Dynamic Velocity Fields

[01.A1.I13, (page 9)]

Aritra GUHA, *AT&T Data Science and AI Research*
Sunrit CHAKRABORTY, *University of Michigan*
Rayleigh LEI, *University of Washington*
XuanLong NGUYEN, *University of Michigan*

Learning and understanding heterogeneous patterns in complex spatio-temporal data is an important and challenging task across domains in science and engineering. In this work we develop a model for learning heterogeneous and dynamic patterns of velocity field data, motivated by applications in the transportation domain. We draw from basic nonparametric Bayesian modeling elements such as infinite hidden Markov model and Gaussian process, and focus on making the learning of such a stochastic model scalable for voluminous and streaming data. This is achieved by employing sequential MAP estimates from the infinite HMM model, an efficient sequential sparse GP posterior computation and refinement of the estimates using Viterbi algorithm, which is shown to work effectively on a careful simulation study. We demonstrate the efficacy of our techniques to the NGSIM dataset of complex multi-vehicle interactions.

100. Change-point detection in high-dimension

[03.A1.I40, (page 25)]

Nilabja GUHA, *University of Massachusetts Lowell*
Jyotishka DATTA,

Many dynamic and random processes in nature go through sudden and significant structural changes. Often the change is in the observable quantity in response to a change in a latent factor. Such 'change-points' are routinely observed across all scientific disciplines and applications, such as economics, epidemiology, social sciences, cybersecurity and finance. Specific examples could be changing regression when the observed variable depends on predictors through a mean structure that changes with time, or change points in data with massive dimensions, such as high-resolution imaging data or complex connected graphs. While there is a substantial literature proposing elaborate methods for detecting change points in different settings, there has been limited consideration of Bayesian methods for change-points that can account for hierarchical models with complex dependence or sparsity structures. This work fills this gap with new statistical tools motivated by real-life applications, by developing a new theoretical framework while retaining efficiency and usefulness in current applications. Here we present our current developments and explore the scope of interdisciplinary applications.

101. A Bayesian Approach to Network Classification

[01.M2.I11, (page 7)]

Sharmistha GUHA, *Texas A&M University*
Abel RODRIGUEZ, *University of Washington*

We propose a novel Bayesian binary classification framework for networks with labeled nodes. Our approach is motivated by applications in brain connectome studies, where the overarching goal is to identify both regions of interest (ROIs) in the brain and connections between ROIs that influence how study subjects are classified. We develop a binary logistic regression framework with the network as the predictor, and model the associated network coefficient using a novel class of global-local network shrinkage priors. We perform a theoretical analysis of a member of this class of priors (which we call the Network Lasso Prior) and show asymptotically correct classification of networks even when the number of network edges grows faster than the sample size. Two representative members from this class of priors, the Network Lasso prior and the Network Horseshoe prior, are implemented using an efficient Markov Chain Monte Carlo algorithm, and empirically evaluated through simulation studies and the analysis of a real brain connectome dataset.

102. Statistical modelling based on distributional representation of wearable data

[04.M2.I59, (page 34)]

Pratim GUHA NIYOGI, *Johns Hopkins Bloomberg School of Public Health*

Vadim ZIPUNNIKOV, *Johns Hopkins Bloomberg School of Public Health*

With advance in wearable technology, each person generates its own distribution which results in samples of distributions. There are multiple ways to characterize distributional-valued observations including probability density function, cumulative distribution function, integrated cumulative function, quantile density, quantile function, and integrated quantile density, and many others. In this talk, we discuss the problem of chosing optimal distributional representation within a linear distributional regression model. Specifically, we construct a J-test to identify the optimal distributional representation. Motivated by Stanford Technology Analytics and Genomics in Sleep Study, we study the associations between stage-specific fragmentation and Insomnia Severity Index. This is a joint work with Dr. Vadim Zipunnikov.

103 . Distributed Inference and Data Compression: A Tale of Two Techniques

[03.A1.I36, (page 24)]

Rajarshi GUHANIYOGI, *Texas A&M University*

TBA

104 . A Bayesian Time-Varying Psychophysiological Interaction (PPI) Model

[01.A2.I19, (page 12)]

Michele GUINDANI, *UCLA, Biostatistics*

TBA

105. Inverses of Matern Covariances on Grids

[01.M2.I10, (page 7)]

Joseph GUINNESS, *Cornell University*

TBA

106 . Covariance estimation with nonnegative partial correlations

[03.E1.I48, (page 29)]

Adityanand GUNTUBOYINA, *University of California Berkeley*
Jake A. SOLOFF, *University of Chicago*
Michael I. JORDAN, *University of California Berkeley*

We study the problem of high-dimensional covariance estimation under the constraint that the partial correlations are nonnegative. The sign constraints dramatically simplify estimation: the Gaussian maximum likelihood estimator is well defined with only two observations regardless of the number of variables. We analyze its performance in the setting where the dimension may be much larger than the sample size. We establish that the estimator is both high-dimensionally consistent and minimax optimal in the symmetrized Stein loss. We also prove a negative result which shows that the sign-constraints can introduce substantial bias for estimating the top eigenvalue of the covariance matrix. This is joint work with Jake A. Soloff (University of Chicago) and Michael I. Jordan (UC Berkeley).

107 . Multi-object Data Integration in the Study of Primary Progressive Aphasia

[03.A1.I36, (page 24)]

Rene GUTIERREZ, *Texas A&M University*

TBA

108 . Bayesian Network Meta-Regression for Aggregate Ordinal Outcomes with Imprecise Categories

[03.M2.I32, (page 21)]

Yeongjin GWON, *University of Nebraska Medical Center*
Ming-Hui CHEN, *University of Connecticut*
Joseph IBRAHIM, *University of North Carolina*

Comparing emerging treatment options is often challenging because of the sparse of direct comparisons from head-to-head trials and inconsistencies in outcome measures among published placebo-controlled trials for each treatment. One potential solution is to aggregate the different outcome measures into a single ordinal response variable for consistent evaluation. The ordinal response variable will inevitably contain unknown response categories because they cannot be directly derived from published data in the literature. In this talk, we propose a statistical methodology to overcome such a common but unresolved issue in the context of network meta-regression for aggregate ordinal outcomes. The proposed approach includes several existing models as special cases and also allows us to conduct a proper statistical analysis in the presence of trials with certain missing categories. We then develop an efficient Markov chain Monte Carlo sam-

pling algorithm to carry out Bayesian computation. A case study demonstrating the usefulness of the proposed methodology is carried out using aggregate ordinal outcome data from 18 clinical trials in treating Crohn's Disease.

109 . Multi-model Ensemble Analysis with Neural Network Gaussian Processes

[01.M2.I10, (page 7)]

Trevor HARRIS, *Texas A&M University*

TBA

110 . Application of Gaussian Mixture Models to Simulated Additive Manufacturing

[Poster Session, (page 16)]

Jason HASSE, *South Dakota State University*
Semhar MICHAEL, *South Dakota State University*
Anamika PRASAD, *Florida International University*

Additive manufacturing (AM) is the process of building components through an iterative process of adding material in specific designs. AM has a wide range of process parameters that influence the quality of the component. This work applies Gaussian mixture models to detect clusters of similar stress values within and across components manufactured with varying process parameters. Further, a mixture of regression models is considered to simultaneously find groups and also fit regression within each group. The results are compared with a previous naive approach.

111. The Kumaraswamy distribution as a failure model under various loss functions and based on progressive censored data

[01.M2.I7, (page 6)]

Amal HELU, *The University of Jordan*

This paper presents the derivation of the maximum likelihood, uniformly minimum variance unbiased, and Bayes estimators for the unknown parameter and the failure function of the Kumaraswamy model based on progressively Type-II censored samples. In particular, the Bayes estimators are obtained using symmetric and asymmetric loss functions with respect to the conjugate prior. A comparison of these estimators is provided, and a real data example is used to demonstrate the proposed methods.

112. Statistics and Data Science at Instacart and Google

[03.M1.I28, (page 19)]

Tim HESTERBERG, *Instacart*

I'll share some stories from work at Instacart and Google, focusing on cases where statistics and data science education has room for improvement. We want to train students for interesting professional lives, not just running t-tests and telling people no, but getting creating in design and analysis and impressing their colleagues.

113. Bootstrapping for Learning Statistics

[Special Session 1, (page 6)]

Tim HESTERBERG, *Instacart*

Statistical concepts such as sampling distributions, standard errors, and P-values are difficult for many students. It is hard to get hands-on experience with these abstract concepts. I think a good way to get that experience is using bootstrapping and permutation tests. I'll demonstrate using a variety of examples.

They're not just for students. I didn't realize just how inaccurate the classical methods are until I started checking them using these methods. Remember that old rule of n >= 30? Try n >= 5000 instead.

The methods are also useful in their own right. We use them all the time at Google – they are easier to use than standard methods (less chance of screwing up), besides being more accurate.

114. Frequent Use of Authentic Assessments in the Statistics Classroom

[03.M1.I28, (page 19)]

Megan HEYMAN, *Rose-Hulman Institute of Technology*

An authentic assessment measures students' ability to apply knowledge of course content in a meaningful way. Students coming from a statistics course are expected to have data analysis skills that are widely applicable, troubleshooting skills that allow them to solve more difficult problems, and notice when something has gone awry. Typical individual assignments often do not assess these valuable skills. This presentation describes how to build authentic assessments of necessary statistical skills directly into your classroom. Learning through authentic assessments during class provides an active and collabo-

rative experience for students. How to provide formative feedback from these assignments often, minimize grading time, and incorporate these in small or large lectures will compose the discussion. Time allowing, we'll also summarize how to use authentic assessments for summative feedback (e.g. a final practicum).

115 . Geostatistical capture-recapture models

[04.M1.I56, (page 33)]

Mevin HOOTEN, *The University of Texas at Austin*

Methods for population estimation and inference have evolved over the past decade to allow for the explicit incorporation of spatial information when using capture-recapture study designs. Traditional approaches to specifying spatial capture-recapture (SCR) models rely on an individual-based detection function that decays as an individual's activity center is farther from a detection location. Traditional SCR models are intuitive because they incorporate mechanisms of animal space use based on their assumptions about activity centers. We generalize SCR models so that they can accommodate a wide range of space use patterns, including for those individuals that may exhibit elliptical utilization distributions. Our approach uses underlying Gaussian processes to characterize the space use of individuals. This allows us to account for multimodal space use patterns as well as nonlinear corridors and barriers to movement. We refer to this class of models as geostatistical capture-recapture (GCR) models. We adapt a recursive computing strategy to fit GCR models to data in stages, some of which can be parallelized. This technique facilitates implementation and leverages modern multicore and distributed computing environments. We demonstrate the application of GCR models by analyzing both simulated data and a data set involving capture histories of snowshoe hares in central Colorado, USA.

116. Achieving Privacy-Utility Balance in Time Series Release Mechanisms Using Multiple Imputation and Stochastic Filtering

[Student Paper Competition 1, (page 5)]

Gaurab HORE, *University of Maryland, Baltimore County*

Ensuring privacy in released data is paramount for data-producing agencies. Privacy and confidentiality issues in data collection and data release mechanisms have seen revolutionary changes in recent years. The procedure guaranteeing desired privacy of released data mostly use the idea of noise addition. While appealing and appropriate for general databases, noise addition for time series data typically changes the sample autocorrelation structure, thereby compromising data utility for time series data. Our previous work, a mechanism called FLIP, tackles the problem of balancing the dual objective of privacy and data utility for time series. However, FLIP requires that the data curator has knowledge or a very good estimator of the spectral density of the original time series, and also the privacy mechanism is built for protecting against attacks that produce a linear prediction of the time series based on the publicly available data and the released time series. In this work, we provide a model-agnostic data release mechanism that ensures privacy through filtering with random coefficients and preserves utility by releasing multiple copies of the time series perturbed using independent random filters. Multiple imputations are popular tools for missing data problems and have been proven effective in data privacy mechanisms as well. In this work, we build time series privacy mechanisms that recommend imputing time series with random filters multiple times in an integrated framework to provide data privacy while preserving up to second-order utility of time series data. Numerical studies explore the practical performance of the new method.

117. Ridge–Type Shrinkage Estimators in Low and High Dimensional Beta Regression Model with Applications

[Special Session 2, (page 18)]

Abdulkadir HUSSEIN, *University of Windsor*
Abdulkadir HUSSEIN, *University of Windsor, Canada*
Reza BELAGHI, *University of Windsor, Canada*
Yasin ASAR, *Necmettin Erbakan University, Konya, Turkey*
AUTHOR 4: S. E. AHMED, BROCK UNIVERSITY, CANADA,

Beta regression (BR) model is useful in the analysis of bounded continuous outcomes such as proportions. It is well-known that for any regression model, the presence of multicollinearity leads to poor performance of the maximum likelihood (ML) estimators. The ridge–type estimators have been proposed to alleviate the adverse effects of the multicollinearity. Furthermore, when some predictors have insignif-

icant or weak effects on the outcomes, it is desired to recover as much information as possible from these predictors instead of discarding them all together. In this talk, we proposed ridge–type shrinkage estimators for low/high dimensional BR model, which address the above two issues simultaneously. We illustrate the methods on two real data sets from econometric and medicine.

118. The Scale Transformed Power Prior for Time-To-Event Data

[03.M1.I26, (page 19)]

Joseph IBRAHIM, *University of North Carolina*
Ethan ALT, *University of North Carolina*
Xinxin CHEN, *University of North Carolina*
Matthew PSIODA, *Glaxo-Smith-Kline (GSK)*
BRADY NIFONG OF GSK IS AN ADDITIONAL CO-AUTHOR,

In clinical trials, data is often available from a previous trial with a different outcome (i.e., binary vs time-to-event). The power prior proposed by Ibrahim and Chen (2000) does not account for different data types in the context discussed here. To accommodate settings in which the historical data and the current data involve different data types, we develop the partial-borrowing scale transformed power prior (straPP) for several commonly used time-to-event models. The partial-borrowing straPP is developed through rescaling the parameter vector from the historical data to align with that of the new data using a transformation based on the Fisher information matrices from the two data models. We also develop the generalized scale transformed power prior (GenstraPP) to provide added robustness for the case in which the scaled parameters are not equal. Several real data sets from the Eastern Cooperative Oncology Group are used to motivate the use of the proposed methods. We demonstrate the advantages of the partial-borrowing straPP over other common priors via simulation and real data analyses using the proportional hazards model and the mixture cure rate model.

119 . Bayesian finite mixture of regression analysis for cancer based on histopathological imaging–environment interactions

[01.E1.I21, (page 13)]

Yunju IM, *University of Nebraska Medical Center*

Cancer is a heterogeneous disease. Finite mixture of regression (FMR)—as an important heterogeneity analysis technique when an outcome variable is present—has been extensively employed in cancer research, revealing important differences in the associations between a cancer outcome/phenotype and covariates. Cancer FMR analysis has been based on clinical, demographic, and omics variables. A relatively recent and alternative source of data comes from histopathological images. Recently, it has been shown that high-dimensional histopathological image features, which are extracted using automated digital image processing pipelines, are effective for modeling cancer outcomes/phenotypes. Histopathological imaging–environment interaction analysis has been further developed to expand the scope of cancer modeling and histopathological imaging-based analysis. Motivated by the significance of cancer FMR analysis and a still strong demand for more effective methods, in this study, we take the natural next step and conduct cancer FMR analysis based on models that incorporate low-dimensional clinical/demographic/environmental variables, high-dimensional imaging features, as well as their interactions. Complementary to many of the existing studies, we develop a Bayesian approach for accommodating high dimensionality, screening out noises, identifying signals, and respecting the "main effects, interactions" variable selection hierarchy. Simulation shows advantageous performance of the proposed approach. The analysis of The Cancer Genome Atlas data on lung squamous cell cancer leads to interesting findings.

120. Multiple Hypothesis Testing To Estimate The Number of Communities in Sparse Stochastic Block Models

[03.A1.I35, (page 23)]

Chetkar JHA, *Washington University in St. Louis*
Li MINGYAO, *Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania*
Ian BARNETT, *Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania*

Network-based clustering methods invariably require the number of communities to be specified *a priori*. Moreover, most of the existing methods for estimating the number of communities assume the number of communities to be fixed and not scale with the network size n. The few methods that assume the number of communities to increase with the network size n are only valid when the average degree d of a network grows at least as fast as

$O(n)$ (i.e., dense case) or lies within a narrow range. This presents a challenge in clustering large scale network data, particularly when the average degree d of a network grows slower than the rate of $O(n)$ (i.e., sparse case). To address this problem, we propose a new sequential procedure utilizing multiple hypothesis tests and the spectral properties of Erdös Rényi graphs for estimating the number of communities in sparse stochastic block models (SBMs). We prove the consistency of our method for sparse SBMs for a broad range of the sparsity parameter. As a consequence, we discover that our method can estimate the number of communities $K^{(n)}\_\star$ with $K^{(n)}\_\star$ increasing at the rate as high as $O(n^{(1-3\gamma)/(4-3\gamma)})$, where $d = O(n^{1-\gamma})$. Moreover, we show that our method can be adapted as a stopping rule in estimating the number of communities in binary tree stochastic block models. We demonstrate the usefulness of our method through numerical simulations and by using it for clustering real single-cell RNA-sequencing datasets.

121. Methane emission detection, localization, and quantification using continuous point-sensors on oil and gas facilities

[Poster Session, (page 16)]

Meng JIA, *Colorado School of Mines*

Reducing methane emissions is a key component of short-term climate action. The oil and gas sector is a promising avenue for methane emission reduction, as it accounts for 22% of global anthropogenic methane emissions and 32% within the U.S. We propose a generic, modular framework for emission event detection (identify the start and end time), localization (localize the emission source), and quantification (estimate emission rate) on oil and gas facilities. The framework uses methane concentration and wind speed and direction data collected by continuous point sensors. The framework is broken up into four steps: 1) background removal and event detection, 2) simulation, 3) localization, and 4) quantification. We evaluate our framework by testing it on a set of 85 controlled releases that vary in duration and size. Our results show that the framework accurately identifies all the events with 82% being localized correctly. 90% of small events (<= 1kg/hr) are quantified within an error range of [-78.1%, 178.6%], while 90% of large events (> 1kg/hr) are quantified within an error range of [-49.6%, 77.4%]. These results demonstrate the effectiveness of our proposed framework in monitoring methane emissions from oil and gas sites.

122. A Random-effects Approach to Regression Involving Many Categorical Predictors and Their Interactions

[04.M1.I52, (page 31)]

Jiming JIANG, *University of California, Davis*
Hanmei SUN, *Shandong Normal University*
Jiangshan ZHANG, *University of California, Davis*

Linear model prediction with a large number of potential predictors is both statistically and computationally challenging. The traditional approaches are largely based on shrinkage selection/estimation methods, which are applicable even when the number of potential predictors is (much) larger than the sample size. A situation of the latter scenario occurs when the candidate predictors involve many binary indicators corresponding to categories of some categorical predictors as well as their interactions. We propose an alternative approach to the shrinkage prediction methods in such a case based on mixed model prediction, which effectively treats combinations of the categorical effects as random effects. We establish theoretical validity of the proposed method, and demonstrate empirically its advantage over the shrinkage methods. We also develop measures of uncertainty for the proposed method and evaluate their performance empirically. A real-data example is considered. This work is joint with Hanmei Sun of Shandong Normal University and Jiangshan Zhang of the University of California, Davis.

123. Development of a Nonstationary Spatio-Temporal Model for Water Vapor Flux Divergence

[01.A2.I17, (page 11)]

Mark KAISER, *Iowa State University*
Jonathan HOBBS, *NASA Jet Propulsion Laboratory, California Institute of Technology*

Developing a spatio-temporal model for a problem that potentially involves nonstationary behavior in both large-scale and small-scale model components requires a systematic examination of distributional forms, scientifically accepted patterns in space and/or time, and the effects of posited changes in small-scale parameters on the ability of the model to represent observed data. We present an example of such model development for an important component of the hydrological cycle in the Midwestern

United States. Our approach is based on Markov random fields and conditional model specification, and we demonstrate the importance of modeling non-constant conditional variances in addition to flexible mean functions in matching higher order properties such as skewness and kurtosis to the observed data.

124 . A Hierarchical Bayesian Entity Resolution Model to Improve Tree Demography Using Overlapping Lidar Scans

[04.M2.I60, (page 34)]

Andee KAPLAN, *Colorado State University*
Lane DREW, *Colorado State University*
Ian BRECKHEIMER, *Rocky Mountain Biological Laboratory*

It has become increasingly common for data to be spread across multiple sources, making it necessary to identify which records refer to the same entity. Entity resolution is the task of estimating this linkage structure in the absence of a unique identifiable attribute. We present a hierarchical Bayesian model for understanding tree growth that incorporates noisy repeat measurements of tree canopy volume and spatial location with entities identified through a spatial entity resolution component. The data comprise overlapping lidar scans of the same forest taken at different points in time. Each scan is post-processed to determine the location of the treetops and an estimate of tree canopy volume. Tree growth functions are then estimated, dependent upon correctly identifying unique individuals across scans without a unique identifying attribute. We identify the unique individuals while quantifying the uncertainty in the linkage process and incorporating it into the inference on tree growth. We discuss the model formulation, including data and process models, and assess performance on both simulated data sets and lidar scans of the Upper Gunnison Watershed.

125 . Prediction Interval for high-dimensional regression with dependent errors

[01.A1.I13, (page 9)]

Sayar KARMAKAR, *University of Florida*

TBA

126 . Current methods for evaluating prediction model performance

[01.A1.I16, (page 10)]

Michael KATTAN, *Department of Quantitative Health Sciences, Cleveland Clinic*

Innovative published prediction models vary tremendously regarding the way their performance is assessed. Numerous metrics and plots have been introduced with little overall guidance for choosing which to report. This problem becomes particularly critical when rival prediction models are being compared, as one metric may favor one model while another metric favors another. This presentation will discuss some straightforward solutions to prediction model assessment and comparison.

127. Inference for Change Points in High Dimensional Mean Shift Models

[01.E1.I24, (page 14)]

Abhishek KAUL, *Washington State University*

We consider the problem of constructing confidence intervals for the locations of change points in a high-dimensional mean shift model. We develop a locally refitted least squares estimator and obtain component-wise and simultaneous rates of estimation of change points. The simultaneous rate is the sharpest available by at least a factor of log p, while the component-wise one is optimal. These results enable existence of limiting distributions. Component-wise distributions are characterized under both vanishing and non-vanishing jump size regimes, while joint distributions of change point estimates are characterized under the latter regime, which also yields asymptotic independence of these estimates. We provide the relationship between these distributions, which allows construction of regime adaptive confidence intervals. All results are established under a high dimensional scaling, in the presence of diverging number of change points. They are illustrated on synthetic data and on sensor measurements from smartphones for activity recognition. (https://arxiv.org/abs/2107.09150).

128 . Opportunities and Challenges in Using External Information in Drug Development

[01.A1.I15, (page 9)]

Amarjot KAUR, *Merck Research Labs*
Bhramori BANERJEE, *Merck*

Rare disease clinical trials, amongst others, have several challenges that make conduct of clinical tri-

als difficult due to small patient populations. There also may also be reasons, including ethical, to not conduct trials using placebo control, where the use of historical control using information on similar patient population can be a meaningful alternative. Usage of external but relevant information, not only reduces the cost of the clinical trial, but also helps increasing the speed of the trial allowing access to treatment in a timely manner. Regulatory agencies are increasingly supportive of innovative methods in design and analysis of clinical trials. While there are many available methods to combine information from historical data including Bayesian framework, each approach has its own advantages and disadvantages. In this presentation we provide user's perspective on commonly used methods to augment external information.

129. Hot off the presses - The SCT Trial of the Year

[01.E1.I22, (page 13)]

Amarjot KAUR, *Merck Research Labs*

TBA

130. MARS via LASSO

[Student Paper Competition 2, (page 7)]

Dohyeong KI, *UC Berkeley*

Multivariate adaptive regression splines (MARS) is a popular method for nonparametric regression introduced by Friedman in 1991. MARS fits simple nonlinear and non-additive functions to regression data. We propose and study a natural lasso variant of the MARS method. Our method is based on least squares estimation over a convex class of functions obtained by considering infinite-dimensional linear combinations of functions in the MARS basis and imposing a variation based complexity constraint. We show that our estimator can be computed via finite-dimensional convex optimization and that it is naturally connected to nonparametric function estimation techniques based on smoothness constraints. Under a few standard design assumptions, we prove that our estimator achieves a rate of convergence that depends only logarithmically on dimension and thus avoids the usual curse of dimensionality to some extent. We implement our method with a cross-validation scheme for the selection of the involved tuning parameter and compare it to the usual MARS method in various simulation and real data settings.

131. Comparison of local powers of some exact tests for a common normal mean with unequal variances

[Special Session 2, (page 18)]

Yehenew KIFLE, *Department of Mathematics and Statistics, University of Maryland Baltimore County (UMBC)*
Bimal SINHA, *Department of Mathematics and Statistics, University of Maryland, Baltimore County (UMBC)*
Alain MOLUH, *Department of Mathematics and Statistics, University of Maryland, Baltimore County (UMBC)*

The inferential problem of drawing inference about a common mean of several independent normal populations with unequal variances has drawn universal attention, and there are many exact and asymptotic tests for testing a null hypothesis 0: = 0 against two-sided alternatives. In this talk, I will provide a review of some of these exact and asymptotic tests, and present theoretical expressions of local powers of the exact tests and a comparison. It turns out that, in the case of equal sample size, a uniform comparison and ordering of the exact tests based on their local power can be carried out even when the variances are unknown.

132. Multiple bias calibration for adjusting selection bias of voluntary samples using data integration

[Special Invited Session 3, (page 12)]

Jae-kwang KIM, *Iowa State University*
Zhonglei WANG, *Xiamen University*
Shu YANG, *North Carolina State University*

Valid statistical inference is challenging when the sample is subject to unknown selection bias. Data integration can be used to correct for selection bias when we have a probability sample from the same population with some common measurements. How to model and estimate the selection probability of the non-probability sample using an independent probability sample is the challenging part of the data integration. We approach this difficult problem by employing multiple candidate models for the propensity score function combined with empirical likelihood. By incorporating multiple propensity score (PS) models into the internal bias calibration constraint in the empirical likelihood setup, the selection bias can be safely eliminated so long as the multiple candidate models contain the true PS model. The bias calibration constraint for the multiple PS model in the empirical likelihood is called the multiple bias

calibration. The multiple PS models can include both missing-at-random and missing-not-at-random models. Asymptotic properties are discussed and some limited simulation studies are presented to compare with the existing methods. The proposed method is applied to simulation platform using the Culture & Community in a Time. of Crisis (CCTC) dataset.

133. Bayesian framework for image analysis in aging studies

[Memorial Session 1, (page 12)]

Namhee KIM, *Rush University Medical Center*
Namhee KIM, *Rush University Medical Center*

This abstract is a tribute to Professor Dalho Kim, who mentored the author during her PhD studies and introduced her to the world of Bayesian statistics. Professor Kim was renowned for his passion for teaching Bayesian statistics and spreading the Bayesian perspective.

The author's interest in Bayesian statistics was sparked when she worked as a full-time statistician for North Gyeongsang Province in South Korea in 1996. She faced a statistical debate on the results of a national income survey between the National Statistical Office and the provincial office, just before the election of a new governor. After discussing the matter with Professor Kim, the author became convinced that the estimates of local areas from the nationwide survey were less reliable. This experience inspired the author to pursue a PhD on a hierarchical Bayesian model to improve small area estimation.

During the memorial session, the author will present her first Bayesian modeling project, which was mentored by Professor Kim. She will also discuss her recent exploration of a new Bayesian framework for image analysis in aging studies.

134 . Modeling massive highly-multivariate nonstationary spatial data with the basis graphical lasso

[01.A2.I17, (page 11)]

William KLEIBER, *University of Colorado Boulder*
Mitchell KROCK, *University of Colorado Boulder*
Dorit HAMMERLING, *Colorado School of Mines*
Stephen BECKER, *University of Colorado Boulder*

We discuss a modeling framework for highly-multivariate spatial processes. This work extends the basis graphical lasso to a multivariate Gaussian process where the basis functions are weighted with Gaussian graphical vectors. We motivate a model where the basis functions represent different levels of resolution and the graphical vectors for each level are assumed to be independent. Using an orthogonal basis grants linear complexity and memory usage in the number of spatial locations, the number of basis functions, and the number of realizations. An additional fusion penalty encourages a parsimonious conditional independence structure in the multilevel graphical model. We illustrate our method on a large climate ensemble from the National Center for Atmospheric Research's Community Atmosphere Model that involves 40 spatial processes.

135. Statistical Analysis of COVID-19 Impacted Bioequivalence Study Data

[03.M1.I29, (page 20)]

Martin KLEIN, *FDA*

This presentation will explore statistical issues caused by missing data in bioequivalence trials impacted by COVID-19. We will examine different scenarios illustrating how missing data can potentially inflate type I error probability or reduce power. We will explore how a multiple imputation approach can be applied in this setting to address the missing data.

136. Piecewise Random-Effects Models for Segmented Longitudinal Trends

[03.A2.I42, (page 26)]

Nidhi KOHLI, *University of Minnesota*

Nonlinear random-effects models for observed, continuous longitudinal data are often used in education, psychology, and broader social sciences to examine individual- and population-level curvilinear development (or growth) over time. The piecewise function is a popular and flexible intrinsically nonlinear function for analyzing segmented trends in individual trajectories over time. Piecewise random-effects models (REMs) allow each segment of the overall developmental trajectory to have a different functional form (e.g., linear-linear, quadratic-linear). The random knot (changepoint), one of the most interesting parameters of the model, is the unknown timepoint of transition from one developmental segment to another. The statistical framework of piecewise REMs has been extended in several ways to enable researchers to analyze different research questions. In the first extension, the piecewise REM allows the detection of multiple knots where data come from a mix-

ture of two or more sub-populations (latent classes). Furthermore, covariates can be incorporated to aid in the identification of latent classes. In the second extension, the model allows for the bivariate modeling of piecewise trajectories for two interdependent longitudinal outcomes (e.g., modeling mathematics and reading achievement scores). In the third and last extension, the model can statistically capture the impact of both dynamic and static group membership on individual outcomes. Each methodological extension is motivated by empirical data applications.

137. Change point detection and localization in a panel of densities

[03.M2.I34, (page 22)]

Piotr KOKOSZKA, *Colorado State University*

Viral load measurements during a prolonged epidemic provide an emerging tool for monitoring its progress. These measurements, aggregated over administrative regions, can be treated as independent scalar observations following a density; one density per day per region. The data can thus be treated as a panel of incompletely observed densities that may change over time, for example, due to the emergence of a new virus variant, change in public health policy, or other changes whose cause may not be obvious at the time of the change. A change point detection problem in such a setting consists of identifying the regions where statistically significant changes have occurred and estimating their time, which can be different for different regions. Challenges to overcome include sparse (or no) observations on certain days in certain regions and the constrained form of random densities, which cannot be treated as unconstrained elements of a Hilbert space. We propose a solution based on an application of Bayes spaces of densities and suitable tools of functional data analysis. Following the presentation of the data and the method, the talk will discuss the theoretical framework.

138 . Learning from Diverse Data in Metric and Preference Learning

[04.M1.I53, (page 32)]

Ramya KORLAKAI VINAYAK, *UW-Madison*
Gokcan TATLI,
Greg CANAL,
Blake MASON,
Rob NOWAK,

Machine learning (ML) algorithms are becoming ubiquitous in various application domains such as public health, genomics, psychology, and social sciences. In these domains the data is obtained from populations that are diverse, e.g., varying demographics, phenotypes, preferences etc. Many ML algorithms focus on learning model parameters that work well on average over the population but do not capture the diversity. On the other hand, such datasets usually have few observations per individual that limits our ability to learn about individuals separately. The question of interest then is, how can we reliably capture the diversity in the data? In this talk, we will address this question in metric and preference learning settings. Learning preferences from human judgements using comparison queries plays a crucial role in cognitive and behavioral psychology, crowdsourcing democracy, surveys in social science applications, and recommendation systems. Models in the literature often focus on learning average preference over the population due to the limitations on the amount of data available per individual. We will discuss some recent results on how we can reliably capture diversity in preferences while pooling together data from individuals.

139. Bayesian image analysis in Fourier space

[01.A2.I19, (page 11)]

John KORNAK, *University of California, San Francisco*
Karl YOUNG, *University of California, San Francisco (Retired)*
Eric FRIEDMAN, *University of California, Berkeley*

For more than 30 years now, Bayesian image analysis has been a leading approach to image reconstruction and enhancement. The idea of the approach is to balance a priori expectations of image characteristics (the prior) with a model for the image degradation process (the likelihood). The conventional Bayesian modeling approach as defined in image space, implements priors that describe inter-dependence between spatial locations on the image lattice (commonly through Markov random field, MRF, models) and can therefore be difficult to model and compute. Bayesian image analysis in Fourier space (BIFS) provides for an alternate approach that can generate a wide range of models, including ones with similar properties to conventional models, but with reduced computational burden; the originally complex high-dimensional estimation problem in image space can

be similarly modeled as a series of (trivially parallelizable) independent one-dimensional problems in Fourier space. A range of prior models that can be formulated in Fourier space will be illustrated, including MRF-matched models and frequency selective models, and these will be compared to conventional models. In addition, extensions will be briefly discussed based on a) a data-driven prior approach and b) transforming to the wavelet domain.

140. On the Variance and Admissibility of Empirical Risk Minimization

[03.E1.I48, (page 29)]

Gil KUR, *MIT*
Alexander RAKHLIN, *MIT*
Eil PUTTERMAN, *Tel Aviv University*

It is well known that Empirical Risk Minimization (ERM) with squared loss may attain minimax sub-optimal error rates. We show that, under mild assumptions, the suboptimality of ERM *must* be due to large bias rather than variance. More precisely, in the bias-variance decomposition of the squared error of the ERM, the variance term necessarily enjoys the minimax rate. In the case of fixed design, we provide an elementary proof of this fact using the probabilistic method. Also, we provide a simple proof to Chatterjee's admissibility theorem – which implies that for any class, ERM cannot be ruled out as an optimal method for some regression function. We then extend the results to the random design setting under mild assumptions.

141 . Case Weighted Adaptive Power Priors for Hybrid External Control Arms

[01.M2.I7, (page 6)]

Evan KWIATKOWSKI, *University of Texas MD Anderson Cancer Center*

We develop method for hybrid analyses that uses external controls from real world data (RWD) to augment internal control arms in randomized controlled trials (RCTs) where the degree of borrowing is evaluated based on similarity between RCT and RWD patients to account for systematic differences (e.g. unmeasured confounders). We develop a novel extension of the power prior where the discounting weight is computed separately for each external control subject based on compatibility with the randomized control data. The discounting weights are assigned using the predictive distribution for the external controls derived via the Bayesian posterior distribution for time-to-event parameters estimated from the RCT. This method is applied in an example based on a completed trial in non-small cell lung cancer using a proportional hazards model with piecewise constant baseline hazard. It is shown that the case weighted adaptive power prior provides robust inference under various forms of heterogeneity in the external control population.

142. Quadratic Prediction Methodology and Calibration of Prediction Intervals Based on Subsampling

[Memorial Session 2, (page 26)]

Soumendranath LAHIRI, *Washington university of St. Louis*

We consider nonlinear prediction of a stationary time series using quadratic functions of the past data. We derive explicit formulae for the best quadratic predictor and its MSPE. We also give conditions under which the quadratic approach improves over the standard linear case and provide a complete characterization for such processes. We next consider the problem of constructing asymptotically valid prediction intervals based on a general point predictor. While much of the existing literature either assumes a parametric time series model or makes specific distributional assumptions (e.g., Gaussian), this work relaxes both and develops a nonparametric method that is applicable to a general stationary sequence. Specifically, we propose a Subsampling method for constructing distribution free prediction intervals for linear and nonlinear prediction methods and establish its validity. For the case of best linear predictor, we also derive the optimal rate of the subsample block size. The results in the prediction context are very nonstandard when compared with the known results on optimal block sizes for the Block Bootstrap/Subsampling in standard variance estimation problems. Finite sample properties of the proposed method are illustrated with simulation.

This is a joint work with Dhrubajyoti Ghosh, Tucker McElroy and Daniel Nordman.

143 . Exploratory subgroup identification in the heterogeneous Cox model: A relatively simple procedure

[03.M1.I26, (page 19)]

Leon LARRY, *Merck*

TBA

144 . Variational Inference: Posterior Threshold Improves Network Clustering Accuracy in Sparse Regimes

[03.E1.I50, (page 29)]

Can LE, *University of California, Davis*
Xuezhen LI, *University of California, Davis*
Can LE, *University of California, Davis*

Variational inference has been widely used in machine learning literature to fit various Bayesian models. In network analysis, this method has been successfully applied to solve the community detection problems. Although these results are promising, their theoretical support is only for relatively dense networks, an assumption that may not hold for real networks. In addition, it has been shown recently that the variational loss surface has many saddle points, which may severely affect its performance, especially when applied to sparse networks. This paper proposes a simple way to improve the variational inference method by hard thresholding the posterior of the community assignment after each iteration. Using a random initialization that correlates with the true community assignment, we show that the proposed method converges and can accurately recover the true community labels, even when the average node degree of the network is bounded. Extensive numerical study further confirms the advantage of the proposed method over the classical variational inference and another state-of-the-art algorithm.

145 . Estimating network-mediated causal effects via spectral embeddings

[03.M1.I30, (page 20)]

Keith LEVIN, *University of Wisconsin-Madison*
Alex HAYES, *University of Wisconsin-Madison*
Mark FREDRICKSON, *University of Michigan*

Causal inference for observational network data is an area of active interest, owing to the ubiquity of network data in the social sciences. Unfortunately, the complicated dependency structure of network data presents an obstacle to many popular causal inference procedures. In this talk, we consider the task of mediation analysis for network data. We present a model in which mediation occurs in a latent node embedding space. Under this model, node-level interventions have causal effects on nodal outcomes, and these effects can be partitioned into a direct effect independent of the network, and an indirect effect, which is induced by homophily. To estimate these network-mediated effects, we embed nodes into a low-dimensional Euclidean space. We then use these embeddings to fit two ordinary least squares models: (1) an outcome model that characterizes how nodal outcomes vary with nodal treatment, controls, and position in latent space; and (2) a mediator model that characterizes how latent positions vary with nodal treatment and controls. We prove that the estimated coefficients are asymptotically normal about the true coefficients under a sub-gamma generalization of the random dot product graph, a widely used latent space model. Further, we show that these coefficients can be used in product-of-coefficients estimators for causal inference. Our method is easy to implement, scales to networks with millions of edges, and can be extended to accommodate a variety of structured data.

146 . Nonprobability follow-up sample analysis: an application to SARS-Cov-2 infection prevalence estimation

[04.M2.I58, (page 33)]

YAN LI, *UNIVERSITY OF MARYLAND AT COLLEGE PARK*
LAURA YEE, *NATIONAL INSTITUTE OF HEALTH*
SALLY HUNSBERGER, *NATIONAL INSTITUTE OF HEALTH*
BARRY GRAUBARD, *NATIONAL INSTITUTE OF HEALTH*

Public health policy makers must make crucial decisions rapidly in some situations such as during a pandemic. Representative health surveys that would aid in decisions may not be feasible due to insufficient time and resources required to enact probability-based sampling that have good response rates. Another approach to designing health surveys is to take a volunteer nonprobability sample from outreach methods such as social media and web surveys. The selected nonprobability sample is then adjusted by propensity-score (PS) pseudoweighting using a "reference" probability-based (representative) survey. The nonprobability sample is followed longitudinally to study cross-sectional changes in measurements of interest in the target population. Nonresponse often occurs in longitudinal surveys. As a result, the samples at later timepoints are subject to both selection bias and nonresponse bias. In this paper, we construct nonresponse-adjusted kernel-weighted pseudoweights (kwNR) for respondents at follow-up vis-

its from a nonprobability sample, then estimate the population mean at each follow-up visit, and develope a Taylor Linearization variance estimator which accounts for variability due to estimating selection propensity and responding propensity. Simulations were conducted to evaluate kernel-weighted vs. nonresponse adjusted kernel-weighted (kwNR) nonprobability sample estimates in terms of bias and variance. We investigate covariate effects on each of the following: baseline survey participation propensity, follow-up responding propensity and the outcome. We applied the proposed kwNR-weighted methods to follow up data from the SARS-Cov-2 seroprevalence survey(ref) study that estimated seroprevalence of SARS-Cov-2. The follow up data was collected the same individuals at six and twelve months post baseline.

147. Bayesian Survival Tree Ensembles with Submodel Shrinkage

[01.A2.I20, (page 12)]

Antonio LINERO, *The University of Texas at Austin*

TBA

148. Challenges of estimating individual treatment effects from clinical data using machine learning

[01.A1.I14, (page 9)]

Ilya LIPKOVICH, *Eli Lilly and Company*
David SVENSSON, *AstraZeneca*
Bohdana RATITCH, *Bayer*
Alex DMITRIENKO, *Mediana*

In this talk we review recent advances in estimating hypothetical individual treatment effects via conditional average treatment effect (CATE), from randomized clinical trials and observational data using statistical/machine learning strategies including S-learning, T-learning, X-learning, R-learning and Causal forests. We discuss common measures for evaluating performance of estimated CATE and illustrate some challenges using simulated data.

149. Conformal Prediction for Network-Assisted Regression

[01.M2.I9, (page 7)]

Robert LUNDE, *Washington University in St. Louis*
Elizaveta LEVINA, *University of Michigan*
Ji ZHU, *University of Michigan*

An important problem in network analysis is predicting a node attribute using nodal covariates and summary statistics computed from the network, such as graph embeddings or local subgraph counts. While standard regression methods may be used for prediction, statistical inference is complicated by the fact that the network summary statistics often exhibit a nonstandard dependence structure. Under a mild joint exchangeability assumption, we show that conformal prediction methods are finite-sample valid for a wide range of network summary statistics. We also prove that a form of asymptotic conditional validity is achievable.

150. How to use causal inference with clinical trial data of CAR-T cell therapies

[01.A1.I14, (page 9)]

Wanying MA, *Novartis Pharmaceuticals Corporation*
Edward WALDRON, *Novartis Pharmaceuticals Corporation*
Julie JONES, *Novartis Pharma AG*

CAR-T cell therapies present a number of new challenges in terms of understanding how multiple complex and related factors influence the manufacture of the product, expansion of the product within the patient's body and clinical response of the patient. We set out to investigate and understand relationships between these factors better with the tool from causal inference.

A crucial component of this work is the use of causal inference techniques to move from simply looking at association between various factors and outcomes to try to understand how these factors are affecting the outcomes, after adjusting the confounders. We also tried to assess whether there are certain clinical meaningful causal pathways. How to properly identify the confounders will also be illustrated.

In this work, we show how DAGs (directed acyclic graphs) help visualize the complex causal relationships within the data and help identify the proper confounders for the targeting causal effects. Moreover, DAGs are a great tool for guiding the discussion over complex data domains. In addition, we also illustrate the usage of causal mediation analysis. Causal mediation analysis breaks down the total effect into the causal mediation effect and direct effect, and it assesses whether the exposure influences the outcome through an intermediate variable, by estimating the average causal mediation effects. It is a powerful tool for explaining a mechanism through which the exposure causally operates to affect the outcome.

This work is part of the collaboration between Novartis and Carnegie Mellon University to understand CAR-T therapy through the lens of principled, flexible, and modern statistical tools. Through this work, we showcase how to utilize the causal inference as a powerful tool in the clinical trial setting.

151 . Exploratory Factor Analysis for Data on a Sphere

[03.A1.I39, (page 25)]

Ranjan MAITRA, *Iowa State University*
Fan DAI, *Michigan Technological University*
Karin DORMAN, *Iowa State University*
Somak DUTTA, *Iowa State University*

Data on high-dimensional spheres arise frequently in many disciplines either naturally or as a consequence of preliminary processing and can have intricate dependence structure that needs to be understood. We develop exploratory factor analysis of the projected normal distribution to explain the variability in such data using a few easily interpreted latent factors. Our methodology provides maximum likelihood estimates through a novel fast alternating expectation profile conditional maximization algorithm. Results on simulation experiments on a wide range of settings are uniformly excellent. Our methodology provides interpretable and insightful results when applied to tweets with the # MeToo hashtag in early December 2018, to time-course functional Magnetic Resonance Images of the average pre-teen brain at rest, to characterize handwritten digits, and to gene expression data from cancerous cells in the Cancer Genome Atlas.

152 . Cumulative Logistic Ordinal Regression with Proportional Odds when the Missing Responses are Nonignorable – Application to Phase III Trial

[01.A1.I15, (page 10)]

Arnab MAITY, *Pfizer*
Huaming TAN, *Pfizer*
Vivek PRADHAN, *Pfizer*
Soutir BANDYOPADHYAY, *Colorado School of Mines*

Missing data are inevitable in clinical trials, and trials that produce categorical ordinal responses are not exempted from this. Typically, missing values in the data occur due to different missing mechanisms, such as missing completely at random, missing at random, and missing not at random. Under a specific missing data regime, when the conditional distribution of the missing data is dependent on the ordinal response variable itself along with other predictor variables, then the missing data mechanism is called nonignorable. In this article we propose an expectation maximization based algorithm for fitting a proportional odds regression model when the missing responses are nonignorable. We report results from an extensive simulation study to illustrate the methodology and its finite sample properties. We also apply the proposed method to a recently completed Phase III psoriasis study using an investigational Pfizer compound.

153 . A Deep Learning Synthetic Likelihood Approximation of a Non-stationary Spatial Model for Extreme Streamflow Forecasting

[01.M1.I4, (page 4)]

Reetam MAJUMDER, *NC State University*
Brian J. REICH, *NC State University*

Extreme streamflow is a key indicator of flood risk, and quantifying the changes in its distribution under non-stationary climate conditions is key to mitigating the impact of flooding events. We propose a non-stationary process mixture model (NPMM) for annual streamflow maxima over the central US (CUS) which uses downscaled climate model precipitation projections to forecast extremal streamflow. Spatial dependence for the model is specified as a convex combination of transformed Gaussian and max-stable processes, indexed by a weight parameter which identifies the asymptotic regime of the process. The weight parameter is modeled as a function of *the annual precipitation for each of the two hydrologic regions within the CUS*, introducing spatio-temporal non-stationarity within the model. The NPMM is flexible with desirable tail dependence properties, but yields an intractable likelihood. To address this, we embed a neural network within a density regression model which is used to learn a synthetic likelihood function using simulations from the NPMM with different parameter settings. Our model is fitted using observational data for 1972-2021, and inference carried out in a Bayesian framework. The two regions within the CUS are estimated to be in different asymptotic regimes based on the posterior distribution of the weight parameter. Annual streamflow maxima estimates based on global climate models for two representative climate pathway scenarios suggest an overall increase in the frequency and magnitude

of extreme streamflow for 2006-2035 compared to the historical period of 1972-2005.

154. TSEC: a framework for online experimentation under experimental constraints

[03.A2.I43, (page 26)]

Simon MAK, *Duke University*

Thompson sampling is a popular algorithm for tackling multi-armed bandit problems, and has been applied in a wide range of applications, from website design to portfolio optimization. In such applications, however, the number of choices (or arms) N can be large, and the data needed to make adaptive decisions require expensive experimentation. One is then faced with the constraint of experimenting on only a small subset of $K \ll N$ arms within each time period, which poses a problem for traditional Thompson sampling. We propose a new Thompson Sampling under Experimental Constraints (TSEC) method, which addresses this so-called "arm budget constraint". TSEC makes use of a Bayesian interaction model with effect hierarchy priors, to model correlations between rewards on different arms. This fitted model is then integrated within Thompson sampling, to jointly identify a good subset of arms for experimentation and to allocate resources over these arms. We demonstrate the effectiveness of TSEC in two applications with arm budget constraints. The first is a simulated website optimization study, where TSEC shows considerable improvements over industry benchmarks. The second is a portfolio optimization application on industry-based exchange-traded funds, where TSEC provides more consistent and greater wealth accumulation over standard investment strategies.

155 . Understanding and approximating leave-one-out cross validation under high-dimensional asymptotics

[03.A2.I45, (page 27)]

Arian MALEKI, *Columbia University*
Arnab AUDDY, *Columbia University*
Haolin ZOU, *Columbia University*
Kamiar RAHNAMA RAD, *City University of New York*

In this talk, we study the problem of parameter tuning or equivalently the problem of out-of-sample risk estimation under the high dimensional settings where standard techniques such as K-fold cross-validation suffer from large biases. Motivated by the low bias of the leave-one-out cross-validation (LO) method, we propose a computationally efficient closed-form approximate leave-one-out formula (ALO) for a large class of regularized estimators. Given the regularized estimate, calculating ALO requires minor computational overhead. With minor assumptions about the data generating process, we obtain a finite-sample upper bound for |LO-ALO|. Our theoretical analysis illustrates that |LO -ALO| converges to zero with overwhelming probability, when both n and p tend to infinity, while the dimension p of the feature vectors may be comparable with or even greater than the number of observations, n. Despite the high-dimensionality of the problem, our theoretical results do not require any sparsity assumption on the vector of regression coefficients. Our extensive numerical experiments show that |LO - ALO| decreases as n and p increase, revealing the excellent finite sample performance of ALO.

156 . Development and evaluation of a computer-aided diagnosis system (CAD) in the absence of a gold standard.

[Special Invited Session 2, (page 10)]

Amita MANATUNGA, *Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University*

We first consider a scenario where functional curves and individual characteristics of subjects are observed and using such subject's information, the decision of the diagnosis is made by a group of experts (in a continuous scale) in the absence of a gold standard. We develop a latent statistical model to predict the diagnosis based on subject's observed data and readings from group of experts and show how this model can be used as computer-aided diagnosis system (CAD) to predict for future subjects, thereby helping physicians to make informed decisions. I will describe our statistical method and its application to a data set. In addition, how to compare the performances of a new CAD to performances of a group of experts still remains as a challenging problem. We propose a new data-driven strategy to aggregate readings from a group of heterogenous experts. Our method is an unsupervised induction method which sequentially assign higher weights to experts who consistently agree with others and to assign lower weights to those who mostly disagree with others, providing a fair evaluation of a new device compared to the consistent opinions among experts.

I describe the method and its application to data and conclude with some remarks.

Joint work with Ying Cui, Qi Yu, Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University and Jeong H. Jang, Yonsei University, South Korea.

157. Nonparametric estimation of the age-at-onset distribution from a cross-sectional sample

[03.E1.I47, (page 28)]

Soutrik MANDAL, *NYU Grossman School of Medicine*
Jing QIN,
Ruth PFEIFFER,

The Washington Ashkenazi Study collected genotype information from enrolled participants to estimate the age-of-onset distribution among BRCA1/2 mutation carriers for breast cancer. Previous studies have only used age-at-disease-onset data from the relatives and genotype information from the probands. Age-of-onset and phenotype of the probands have never been used due to two potential sources of bias that occur in all cross-sectional studies or prevalent surveys: first, selection bias, as individuals had to be alive at the time of the study in order to be enrolled, and second, survival bias in prevalent cases, as those previously diagnosed with breast cancer had to survive long enough from disease onset to study enrollment. We propose a simple and innovative nonparametric approach that accommodates these biases to estimate the age-of-onset distribution for a disease from a cross-sectional sample of the population that includes individuals with prevalent disease. First, we estimate the joint distribution of two event times, the age-of-onset of breast cancer and the time to death after breast cancer onset. We accommodate the survival bias in the cross-sectionally sampled probands by conditioning on their survival until the age at study enrollment. From these joint probabilities we obtain unbiased estimates of the age-at-onset distribution of disease using a computationally efficient expectation-maximization (EM) algorithm, and also a fully efficient estimate of the mutation prevalence in the source population. We demonstrate the performance of our method using simulation studies under different levels of truncation, and finally, analyze the data from female participants in the Washington Ashkenazi Study to estimate the age-specific penetrance for breast cancer.

158. Computational Lenses in learning combinatorial structures in statistics and machine learning.

[Special Invited Session 5, (page 21)]

Rahul MAZUMDER, *MIT Sloan School of Management*

Optimization problems involving a discrete or combinatorial structure frequently arise in statistics and machine learning: best subset selection, classification, robust statistics, and decision trees, among others. I will discuss how such combinatorial problems have inspired new recent developments in mathematical optimization and how these algorithms can inform our computational and statistical understanding of these problems. For example, I will discuss how tools in convex and discrete optimization can be used to solve, to global optimality, instances of the best subset selection problem with one million features and thousands of samples. Time permitting, I will also discuss approaches for learning decision trees and shuffled regression.

159. Global-local Shrinkage Priors for Multimodal Data Integration

[01.M1.I6, (page 5)]

Omar MELIKECHI, *Duke University*

TBA

160. Constrained Spline Density Estimation and Applications

[Special Invited Session 1, (page 8)]

Mary MEYER, *Colorado State University*
Hanxiao JING, *Colorado State University*
Xin CHEN,

For estimating a regression function, splines are generally preferred over kernel smoothers. One advantage to using spline is that shape constraints such as monotonicity or convexity are easily imposed with splines. Kernel methods are a popular choice for density estimation, but there is again a good argument that penalized splines are superior. Constraints such as unimodal or decreasing are more readily implemented with splines compared to kernels, and the spline density estimates enjoy a faster convergence rate. In this talk, we review constrained spline density estimation and look at some applications including robust regression, quantile regression, estimating the deconvolution density, and testing for sampling bias. Each of these uses the penalized, shape-constrained spline density estimator.

161 . Modeling heterogeneity in hierarchically structured data for source identification problems

[Special Session 2, (page 18)]

Semhar MICHAEL, *South Dakota State University*
Andrew SIMPSON, *South Dakota State University*
Dylan BORCHERT, *South Dakota State University*
Christopher SAUNDERS, *South Dakota State University*
COAUTHOR 4 - LARRY TANG, LIANSHENG.TANG@UCF.EDU, UNIVERSITY OF CENTRAL FLORIDA,

Finite mixtures are known for modeling heterogeneity in data. The Gaussian mixture model is the most commonly used by practitioners. The common way of estimating parameters of this model assumes that the data sampled through a simple random sampling process. However, in some applications such as the forensic source identification problem, data has hierarchical structure in addition to the heterogeneity. In this work, we will focus on identifying and characterizing subpopulations in a population when there are hierarchically structured data. This will be done through semi-supervised finite mixture models that are adjusted for the hierarchical sampling procedure. We will illustrate this based on a simulation study using synthetic data and a classical glass datasets.

162 . Statistical models for mixed frequency data and their applications in forecasting economic indicators

[Special Invited Session 5, (page 21)]

George MICHAILIDIS, *UCLA*

We discuss the problem of modeling and analysis of time series data that evolve at different frequencies (e.g., quarterly-monthly). We present univariate and multivariate modeling paradigms, outline and address technical challenges and illustrate their performance in forecasting tasks involving key macroeconomic indicators.

163. Estimating the Selected Treatment Effect Using a Two-Stage Adaptive Design

[04.M2.I59, (page 34)]

Neeraj MISRA, *Indian Institute of Technology, Kanpur, India*
Masihuddin MASIHUDDIN, *Indian Institute of Technology, Kanpur, India*

Adaptive designs are commonly used in clinical and drug development studies for optimum utilization of available resources. In this work, we consider the problem of estimating the effect of the selected (better) treatment using a two-stage adaptive design, particularly called as the "Drop-the-Losers Design (DLD)". We consider two treatments whose effects are described by independent Gaussian distributions having different unknown means and a common known variance. The effective (or better) treatment is adjudged based on numerical values of some statistic corresponding to the two treatments. To select the more effective treatment, the two treatments are independently administered to n_1 subjects each and the treatment corresponding to the larger sample mean is selected. To study the effect of the adjudged more effective treatment (i.e. estimating its mean), we consider the two-stage DLD in which n_2 subjects are further administered the adjudged more effective treatment in the second stage of the design. We obtain some admissibility and minimaxity results for estimating the mean effect of the adjudged more effective treatment. The maximum likelihood estimator (a weighted average based on the first stage and the second stage sample means) is shown to be minimax and admissible. We show that the uniformly minimum variance conditionally unbiased estimator (UMVCUE) of the selected treatment proposed by Bowden and Glimm (2008) is inadmissible and obtain an improved estimator. In the process, we also derive a sufficient condition for inadmissibility of an arbitrary location and permutation equivariant estimator and provide dominating estimators in cases, where this sufficient condition is satisfied. For the case of unknown variance σ^2, we also derive the UMVCUE of the selected treatment effect. The minimaxity and admissibility of the maximum likelihood estimator is also shown, even in the case of unknown variance σ^2. The mean squared error and the bias performances of various competing estimators are compared via a simulation study.

References

Bowden, J. and Glimm, E. (2008). Unbiased estimation of selected treatment means in two-stage trials. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 50(4):515-527.

164 . Observing Relations in Mixed Corpora through Abstractive Summarization with Dimension Reduction: A

Topic Modeling Approach

[03.A2.I43, (page 27)]

R. Cole MOLLOY, *JHU/APL*

Observing Relations in mixed Corpora through Abstractive Summarization (ORCAS) is a software tool designed to ingest multiple unstructured text sources and produce a short abstractive summary containing the key information across multiple sources. The algorithmic underpinning of ORCAS is a transformer-based encoder-decoder language model which performs well on datasets of moderate length but cannot ingest long texts. In this talk, we discuss the effectiveness of ORCAS on various corpora and our work to overcome its length limitations via the addition of a dimension-reduction step to the analytic pipeline. In particular, we cover the application of topic modeling as a technique for reducing the size of text data while preserving its distributional characteristics, and the effectiveness of incorporating topic modeling into the ORCAS summarization tool.

165. Bayesian estimation of local volatility from option pricing data

[04.M2.I60, (page 34)]

Anirban MONDAL, *Case Western Reserve University*
Kai YIN, *Case Western Reserve University*

Local volatility is an important quantity in option pricing, portfolio hedging, and risk management. It is not directly observable from the market; hence calibrations of local volatility models are necessary using observable market data. Unlike most existing point-estimate methods, we cast the large-scale nonlinear inverse problem into the Bayesian framework, yielding a posterior distribution of the local volatility, which naturally quantifies its uncertainty. This extra uncertainty information enables traders and risk managers to make better decisions. To alleviate the computational cost, we apply Karhunen–Lòeve expansion to reduce the dimensionality of the Gaussian Process prior for local volatility. A modified two-stage adaptive Metropolis algorithm is used to sample the posterior probability distribution, which further reduces computational burdens caused by repetitive numerical forward option pricing model solver and time of heuristic tuning. We demonstrate our methodology with both synthetic and market data.

166. Estimating the fraction of anomaly points

[04.M1.I53, (page 32)]

Debashis MONDAL, *Washington University in St Louis*

In the past decade, there has been much progress and growth of anomaly detection algorithms in machine learning literature with applications ranging from insider threat detection, bio-surveillance, computer security, data cleaning and scientific discovery. In this talk, I will address the question of estimating the fraction of anomaly points in high-dimensional data. The focus will be on a semi-supervised learning setting in which we have a clean sample at training time and a contaminated sample at classification time. I will discuss algorithms that assign low dimensional scores to data points and semi-parametric mixture models that can be applied to estimate the fraction of anomaly points. The work arose in collaboration with former PhD student Si Liu and EECS professor Tom Dietterich.

167. Semiparametric adaptive estimation under informative sampling

[01.M1.I3, (page 4)]

Kosuke MORIKAWA, *Osaka University*
Jae Kwang KIM, *Iowa State University*
Yoshikazu TERADA, *Osaka University*

In probability sampling, sampling weights are often used to remove the selection bias in the sample. The Horvitz-Thompson estimator is well-known to be consistent and asymptotically normally distributed; however, it is not necessarily efficient. This study derives the semiparametric efficiency bound for various target parameters by considering the survey weights as random variables and consequently proposes two semiparametric estimators with working models on the survey weights. One estimator assumes a reasonable parametric working model, but the other estimator requires no specific working models by using the debiased/double machine learning method. The proposed estimators are consistent, asymptotically normal, and can be efficient in a class of regular and asymptotically linear estimators. A limited simulation study is conducted to investigate the finite sample performance of the proposed method. The proposed method is applied to the 1999 Canadian Workplace and Employee Survey data.

168. Deep Neural Networks for Nonparametric Interaction Models with Diverging Dimension

[04.M2.I61, (page 34)]

Debarghya MUKHERJEE, *Princeton University*
Sohom BHATTACHARYA, *Princeton University*
Jianqing FAN, *Princeton University*

Deep neural networks have achieved tremendous success due to their representation power and adaptation to low-dimensional structures. Their potential for estimating structured regression functions has been recently established in the literature. However, most of the studies require the input dimension to be fixed and consequently ignore the effect of dimension on the rate of convergence and hamper their applications to modern big data with high dimensionality. In this talk, we present our theoretical findings for the analysis of a k^{th} order nonparametric interaction model in both growing dimension scenarios (d grows with n but at a slower rate) and in high dimension ($d > n$). In the latter case, sparsity assumptions and associated regularization are required to obtain optimal convergence rates. A new challenge in diverging dimension setting is in calculation mean-square error, the covariance terms among estimated additive components are an order of magnitude larger than those of the variances and can deteriorate statistical properties without proper care. We introduce a critical debiasing technique to amend the problem. We show that under certain standard assumptions, debiased deep neural networks achieve a minimax optimal rate both in terms of (n, d). Our proof techniques rely crucially on a novel debiasing technique that makes the covariances of additive components negligible in the mean-square error calculation. In addition, we establish the matching lower bounds.

169. A new central limit theorem for the augmented IPW estimator: variance inflation, cross-fit covariance and beyond

[01.M1.I2, (page 3)]

Rajarshi MUKHERJEE, *Harvard T.H. Chan School of Public Health*
Kuanhao JIANG, *Harvard University*
Subhabrata SEN, *Harvard University*
Pragya SUR, *Harvard University*

In recent times, inference for the ATE in the presence of high-dimensional covariates has been extensively studied. Among the diverse approaches that have been proposed, augmented inverse propensity weighting (AIPW) with cross-fitting has emerged as a popular choice in practice. In this work, we study this cross-fit AIPW estimator under well-specified outcome regression and propensity score models in a high-dimensional regime where the number of features and samples are both large and comparable. Under assumptions on the covariate distribution, we establish a new CLT for the suitably scaled cross-fit AIPW that applies without any sparsity assumptions on the underlying high-dimensional parameters. Our CLT uncovers two crucial phenomena among others: (i) the AIPW exhibits substantial variance inflation that can be precisely quantified in terms of the signal-to-noise ratio and other problem parameters, (ii) the asymptotic covariance between the pre-cross-fit estimates is non-negligible even on the root-n scale. In fact, these cross-covariances turn out to be negative in our setting. These findings are strikingly different from their classical counterparts. On the technical front, our work utilizes a novel interplay between three distinct tools—approximate message-passing theory, the theory of deterministic equivalents, and the leave-one-out approach. We believe our proof techniques should be useful for analyzing other two-stage estimators in this high-dimensional regime. Finally, we complement our theoretical results with simulations that demonstrate both the finite sample efficacy of our CLT and its robustness to our assumptions.

170. Bayesian Predictive Inference for Small Areas Using a Non-Probability Sample with Supplemental Information

[Memorial Session 1, (page 12)]

Balgobin NANDRAM, *Professor*

We show how to use supplemental information from a small probability sample (ps) to do Bayesian predictive inference for finite population means of small areas using a relatively larger non-probability sample (nps). We focus on the most practical situation when there are common covariates; the nps has the study variable but no survey weights and the ps has survey weights but no study variable. We assume that the population model is correct and any functional relation between the study variable and the covariates is unspecified. Data preparation is necessary, and there are three steps, which are a double mass imputation, stratification of the population (not the samples), and creating a spatial structure

to accommodate the covariates.Our main Bayesian analysis uses the conditional auto-regressive model, which helps to accommodate the covariates without incorporating them into the model, thereby avoiding a functional relation between the study variable and the covariates. However, the actual small areas are not part of the model, but we need to keep track of them, and the strata are modeled as the "small areas". Our procedure allows a small area (not a stratum) to participate in several strata, and this helps to mitigate over-shrinkage, which is common in small area models. Using an illustrative example on body mass index data, our method appears to work well when compared to a standard method with a linear regression of the study variable on the covariates.

171. On Estimation of the Logarithm of the Mean Squared Prediction Error of A Mixed-effect Predictor

[04.M1.I52, (page 31)]

Thuan NGUYEN, *Oregon Health and Science University*
Jianling WANG, *Shandong University*
Yihui LUAN, *Shandong University*
Jiming JIANG, *University of California, Davis*

The mean squared prediction error (MSPE) has been used as an important measure of uncertainty in small area estimation. It is desirable to produce a second-order unbiased MSPE estimator. The task is difficult, however, especially if one needs to take into consideration that an MSPE estimator needs to be positive, or at least nonnegative. In fact, very few MSPE estimators have the property of being both second-order unbiased and guaranteed positive. We consider an alternative, easier approach of estimating the logarithm of the MSPE (log-MSPE), which avoids the issue of positivity. A second-order unbiased estimator of the log-MSPE is derived using the Prasad-Rao linearization method. Empirical studies demonstrate superiority of the proposed log-MSPE estimator over a naive log-MSPE estimator as well as an existing method known as McJack. A real-data example is considered.

172. A Hybrid Empirical Likelihood for Time Series

[03.M2.I34, (page 22)]

Dan NORDMAN, *Iowa State University*
Haihan YU, *Iowa State University*
Mark KAISER, *Iowa State University*

Frequency domain analysis of time series is often complicated by periodogram-based statistics having complex variances, so that approximations from resampling or empirical likelihood (EL) can be helpful. Existing versions of periodogram-based EL for time series, though, are restricted to linear processes and special spectral parameters. This talk introduces a new spectral EL (SEL) method by merging two different EL frameworks for time series, namely, block-based and periodogram-based EL. The resulting SEL statistics have some nice features for inference: these admit chi-square limits under mild conditions and can be coupled to an effective bootstrap procedure. The scope of EL for time series inference is then greatly expanded as SEL: can handle any spectral parameters; is valid for general processes (including nonlinear); and has a provable bootstrap that provides a novel alternative to other resampling plans in the frequency domain. Numerical studies suggest that the method has good performance, which is also demonstrated with a real example.

173. Fast methods for conditional simulation, the key to spatial inference.

[Plenary Lecture 1, (page 15)]

Doug NYCHKA, *Colorado School of Mines*

An advantage of a Gaussian process (GP) model for surface fitting is the companion estimates of the function's uncertainty. The standard method for assessing uncertainty of a GP estimate is through conditional simulation, a Monte Carlo sampling algorithm of the multivariate Gaussian distribution. Conditional simulation is a powerful tool, for example allowing for Monte Carlo based uncertainty on surface contours (level sets), a difficult and nonlinear inference problem. This algorithm, however, has two bottlenecks: generating spatial predictions on large, but regular grids and also simulation of a Gaussian process on both a large regular grid and at irregular locations. Accurate approximations are proposed that allow for fast computation of both these steps. The computational efficiency is achieved by relying on the fast Fourier transform for 2D convolution and also sparse matrix multiplication. Under common spatial applications a speedup by a factor from 10 to a 100 or more is obtained and makes it possible to determine uncertainty of GP estimates on a laptop and in often an interactive setting. Besides the practical benefits of this speedup their accuracy are examples of the â€œscreening effectâ€ for spatial prediction and are related to the errors bounds in interpolation when the

GP is related to an element in a reproducing kernel Hilbert space.

See Bailey, Maggie D., Soutir Bandyopadhyay, and Douglas Nychka. "Adapting conditional simulation using circulant embedding for irregularly spaced spatial data" Stat 11.1 (2022): e446.

174. Solving Drug Development Challenges with Data Science: Combining Statistical Inference, Modeling, and Machine Learning for Effective Decision Making

[Special Invited Session 2, (page 10)]

David OHLSSEN, *Novartis*

This presentation examines three data science approaches for solving drug development challenges. The first approach uses classical statistical thinking, applying statistical inference and experimental design to provide solutions with well-defined operating characteristics. The second approach relies on modeling to generate a good approximation, then utilizes simulation to propagate uncertainty and serve as a basis for decision making. The third approach employs machine learning to simplify complex high-dimensional problems and provide predictive solutions. We illustrate the value of these approaches with a case study from a psoriatic arthritis drug development program, which highlights the richness and complexity of data collected during clinical development. Although each approach can provide effective solutions to specific problems, combining all three approaches is often necessary to find the best solution. This is especially the case when multiple data domains and sources are used to solve a problem. We examine a class of problems where it's not possible to conduct randomized experiments to compare two treatments or options. In such cases, we explore solutions that combine statistical inference, machine learning, and modeling. We also examine the challenge of identifying prognostic and predictive factors using the knockoff approach, which blends control of operating characteristics with complex modeling. This presentation demonstrates how data science approaches can effectively address drug development challenges and support decision making.

175. Fitting Classification Trees to Complex Survey Data

[Special Invited Session 3, (page 13)]

Jean OPSOMER, *Westat*
Minsun RIDDLES, *Westat*

Classification tree algorithms are a convenient method to perform variable selection and obtain interpretable structures relating covariates and an outcome of interest. When fitting classification trees to survey data, it is common to ignore sampling weights as well other design characteristics such as stratification and clustering. However, unless the survey design is uninformative, there is a risk that the inference for the classification tree is incorrect. A particular application in which this is a concern is the construction of nonresponse adjustment cells, a key step in the development of survey weights. We propose an extension of the popular Chi-square Automatic Interaction Detector (CHAID) approach that accounts for the design by applying a Rao-Scott correction in its classification criterion. We discuss the statistical properties of the resulting algorithm under a design-based framework. We compare its performance to existing weighted and unweighted algorithms, and we illustrate the use of the method using data from the U.S. American Community Survey.

176. Optimal Bayesian Inference for High-dimensional Linear Regression Based on Sparse Projection-posterior

[Student Paper Competition 2, (page 7)]

Samhita PAL, *North Carolina State University*

We consider a novel Bayesian approach to estimation, uncertainty quantification, and variable selection for a high-dimensional linear regression model under sparsity. The number of predictors may even be nearly exponentially large relative to the sample size. Instead of traditionally putting a spike-and-slab prior or its variant continuous shrinkage prior on the regression coefficients, we put a conjugate normal prior initially disregarding sparsity. To make an inference, instead of the original multivariate normal posterior, we use the posterior distribution induced by a map transforming the vector of regression coefficients to a sparse vector obtained by minimizing the sum of squares of deviations plus a suitably scaled l1-penalty on the vector. We show that the resulting sparse projection-posterior distribution concentrates around the true value of the parameter at the optimal rate adapted to the sparsity of the vector. We obtain a key sign-consistency result that shows that the true sparsity structure gets a large sparse projection-posterior probability, thereby consistently selecting the active predictors. We further show that an appropriately re-centred credible ball has the correct asymptotic frequentist coverage. Finally, we describe

how the computational burden of can be distributed to many machines, each dealing with only a small fraction of the whole dataset. We conducted a comprehensive simulation study under a variety of settings and found that the proposed method performs well for finite sample sizes.

177. Survival Bandits

[04.M1.I54, (page 32)]

Yinghao PAN, *University of North Carolina at Charlotte*
Eric LABER, *Duke University*
Yingqi ZHAO, *Fred Hutchinson Cancer Research Center*

We consider a contextual survival bandit setting, a variant of the classical multi-armed bandit problem in which the reward for each individual is a survival time subject to right censoring. First, we design a Thompson sampling algorithm that randomly allocates individuals to treatments in an adaptive manner based on Bayesian posterior distributions. Next, we propose a weighted M-estimator for constructing valid confidence regions using data collected from the Thompson sampling algorithm mentioned above. Asymptotic properties of the proposed weighted M-estimator are established by careful use of martingale theory.

178. Finite mixture of regression models for censored data based on the skew-t distribution

[Poster Session, (page 16)]

Jiwon PARK, *University of Connecticut*

Finite mixture models have been widely used to model and analyze data from heterogeneous populations. Moreover, data of this kind can be subject to upper and/or lower detection limits because of the restriction of experimental apparatuses. Another complication arises when measures of each mixture component depart significantly from normality, for instance, asymmetry and fat tails behavior, simultaneously. For such data structures, we propose a robust model for censored data based on finite mixtures of skew-t distributions. We develop an analytically simple expectation conditional maximization (ECM) algorithm to estimate the model parameters by monotonically maximizing the observed data log-likelihood. The algorithm has closed-form expressions at the E-step that rely on formulas for the mean and variance of truncated skew-t distributions. Furthermore, a general information-based method for approximating the asymptotic covariance matrix of the estimators is also presented. Results obtained from the analysis of both simulated and real datasets are reported to demonstrate the effectiveness of the proposed method.

179 . Title: Adaptive finite element type decomposition of Gaussian random fields

[03.A1.I40, (page 25)]

Debdeep PATI, *Texas A&M University*
Jaehoan KIM, *Texas A&M University*
Anirban BHATTACHARYA, *Texas A&M University*

In this talk, we investigate a general class of approximate Gaussian processes (GP) obtained by taking linear combination of compactly supported basis functions with the basis coefficients endowed with a sparse dependence structure. This general class includes two highly scalable approximate GP methods: the finite element approximation of the stochastic partial differential equation associated with Matern GP and a linear approximation of a general GP on a regular lattice. We propose prior distributions for the number of basis functions to yield optimal rate of posterior convergence of the underlying function, adaptively over a large class of smooth functions. We also provide two scalable algorithms and numerics to illustrate the methodology.

180 . Use of longitudinal serum markers as early predictors of treatment outcome in germ cell cancers

[01.A1.I16, (page 10)]

Sujata PATIL, *Cleveland Clinic*

Many germ cell cancers secret alpha-fetoprotein (AFP) and human chorionic gonadotrophin (hCG). The decline of these two serum markers help determine whether the primary tumor has regressed. These markers are gathered serially to indicate whether a patient's tumor is responding to a treatment regimen. The statistical issues surrounding how to use this serially collected data to predict if a patient is responding to therapy, and whether an early change in therapy is warranted have not been clearly outlined and is the primary purpose of this work. Herein, we describe and compare commonly used methods to define marker decline in germ cell cancers as predictors of treatment response. We also propose an analytic framework that incorporates subject-level correlation and intrinsic disease and patient charac-

teristics into the prediction. Sample size implications to studying marker decline as an indication for early treatment change in a prospective study discussed.

181 . Means as outcomes: Improving interpretation when hazards are non-proportional.

[01.M2.I8, (page 6)]

Mitchell PAUKNER, *Northwestern University*

The field of clinical trials has experienced a level of innovation that has forced statisticians to re-evaluate how to estimate, test, and communicate treatment effects. In an era where cutting-edge treatment methods consistently and predictably produce non-proportional hazards (NPH), it is important to establish a new precedent for the design and analysis of clinical trials that have time-to-event outcomes as primary endpoints.

In this talk, we discuss the use of mean-based methodology, mainly restricted mean survival time (RMST) and window mean survival time (WMST) as viable alternatives to the classic logrank test and Cox model, especially when the proportional hazards assumption fails with advantages that range from improved power, flexibility in design choices, and coherent interpretation under any survival distribution.

182. Simple Binary Hypothesis Testing: Locally-Private and Communication-Efficient

[04.M1.I53, (page 32)]

Ankit PENSIA, *IBM Research*
Amir ASADI, *University of Cambridge*
Varun JOG, *University of Cambridge*
Po-Ling LOH, *University of Cambridge*

Simple binary hypothesis testing is a fundamental problem in statistics and it is well-known that its sample complexity is characterized by the Hellinger divergence between the two candidate distributions. In this talk, we discuss the problem of simple binary hypothesis testing under communication constraints and local differential privacy (LDP) constraints, wherein each sample is privately mapped to a message from a finite set of messages before being revealed to the statistician. These constraints are common in distributed estimation tasks.

In this talk, we will present two kinds of results: 1. (Statistical) A comprehensive characterization of the minmax optimal sample complexities for all parameter ranges of these constraints, whereas prior work had focused only on the high-privacy regime (epsilon < 1 in LDP). 2. (Computational) The first polynomial-time algorithms with near-optimal performance under these constraints, whereas prior algorithms were either exponential-time or sub-optimal.

183. Covariate-adaptive randomization inference in matched designs

[01.M1.I2, (page 3)]

Sam PIMENTEL, *UC Berkeley*

It is common to conduct causal inference in matched observational studies by proceeding as though treatment assignments within matched sets are assigned uniformly at random and using this distribution as the basis for inference. This approach ignores observed discrepancies in matched sets that may be consequential for the distribution of treatment, which are succinctly captured by within-set differences in the propensity score. We address this problem via covariate-adaptive randomization inference, which modifies the permutation probabilities to vary with estimated propensity score discrepancies and avoids requirements to exclude matched pairs or model an outcome variable. We show that the test can achieve type I error control arbitrarily close to the nominal level when large samples are available for propensity score estimation. We characterize the large-sample behavior of the new randomization test for a difference-in-means estimator of a constant additive effect. We also show that existing methods of sensitivity analysis generalize effectively to covariate-adaptive randomization inference. Finally, we evaluate the empirical value of covariate-adaptive randomization procedures via comparisons to traditional uniform inference in matched designs with and without propensity score calipers and regression adjustment using simulations and an analysis of genetic damage among welders.

184. A Matrix Ensemble Kalman Filter-based Multi-arm Neural Network to Adequately Approximate Deep Neural Networks

[Poster Session, (page 17)]

Ved PIYUSH, *Department of Statistics, University of Nebraska - Lincoln*

Deep Learners (DLs) are the state-of-art predictive mechanism with applications in many fields

requiring complex high dimensional data processing. Although conventional DLs get trained via gradient descent with back-propagation, Kalman Filter (KF)-based techniques that do not need gradient computation have been developed to approximate DLs. We propose a multi-arm extension of a KF-based DL approximator that can mimic DL when the sample size is too small to train a multi-arm DL. The proposed Matrix Ensemble Kalman Filter based multi-arm ANN (MEnKF- ANN) also performs explicit model stacking that becomes relevant when the training sample has an unequal-size feature set. Our proposed technique can approximate Long Short-term Memory (LSTM) Networks and attach uncertainty to the predictions obtained from these LSTMs with desirable coverage. We demonstrate how MEnKF-ANN can "adequately" approximate an LSTM network trained to classify what carbohydrate substrates are digested and utilized by a microbiome sample whose genomic sequences consist of polysaccharide utilization loci (PULs) and their encoded genes. The scripts to reproduce the results in this paper are available at https://github.com/VedPiyush/MEnKF-ANN-PUL.

185 . On Estimation of Function-on-function Regression Coefficients with Brownian Berkson Errors

[03.M2.I34, (page 22)]

Paramahansa PRAMANIK, *University of South Alabama*

In this paper, we introduce a new methodology to determine an optimal coefficient of function-on-function regression with a stochastic differential Berkson equation. We assume the response variable, unobserved true predictor, the best available observed measure of the true predictor and the regression coefficients are functions of time and error dynamics following a stochastic differential equation. First, we construct an objective function as a time-dependent Mean Square Error (MSE) and then minimize it with respect to regression coefficients subject to stochastic Berkson error dynamics. A Feynman-type path integral approach is used to determine a Schrodinger-type equation that has the complete information of the system. Using first-order conditions for these coefficients gives us a closed-form solution.

186. Country-specific Estimation of Verbal Autopsy Misclassification in Improving Global Mortality Surveillance

[03.E1.I47, (page 28)]

Sandipan PRAMANIK, *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health*
Abhirup DATTA, *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health*
Scott ZEGER, *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health*

Cause-specific mortality fractions (CSMFs), the proportion of deaths in a population attributable to a given cause, are routinely monitored to understand disease trends and inform public health policies. For this purpose, verbal autopsy (VA) and computer-coded VA algorithms (CCVAs) are widely used for collecting individual-level cause of death (COD) data; particularly, in many low-and-middle-income countries (LMICS) where a full autopsy often cannot be performed. Despite their increased acceptance due to low cost, non-invasive nature, and scalability, CCVAs systematically misdiagnose the cause of many deaths, which leads to biased raw CSMF estimates. Recent research on VA-calibration provides a global estimate of CCVA misclassification rates for all countries by leveraging a limited labeled dataset of VA records paired with a more clinically informed COD diagnosis like a minimally invasive tissue sampling (MITS). When calibrated, this produces CSMF estimates with increased accuracy and validity. In this research, we take advantage of a small number of MITS-VA COD pairs from the Child Health and Mortality Prevention Surveillance (CHAMPS) Network, an ongoing mortality surveillance project, and investigate heterogeneity in misclassification across countries. In the presence of heterogeneity, for CCVA output with COD determined as the top cause we propose a nested framework of Bayesian hierarchical methods for a small-sized labeled dataset to facilitate statistical inference on misclassification rates. We introduce effect sizes with clear interpretations that account for different sources of heterogeneity and allow the framework to learn the degree of heterogeneity from the data. This (i) addresses local heterogeneity in the misclassification rates, (ii) accounts for a global trend in the systematic bias of CCVA algorithms towards specific causes, and (iii) produces country-specific misclassification estimates via Bayesian transfer learning irrespective of their availability of labeled data. This generalizes the existing implementation of VA-calibration and complements the ongoing effort of leveraging the large VA

database in improving global mortality surveillance.

187. Global-Local Shrinkage Priors for Asymptotic Point and Interval Estimation of Normal Means under Sparsity

[Student Paper Competition 2, (page 7)]

Zikun QIN, *University of Florida*

The paper addresses asymptotic estimation of normal means under sparsity. The pri- mary focus is estimation of multivariate normal means where we obtain exact asymptotic minimax error under global-local shrinkage prior. This extends the corresponding uni- variate work of Ghosh and Chakrabarti (2017). In addition, we obtain similar results for the Dirichlet-Laplace prior as considered in Bhattacharya, Pati, Pillai, and Dun- son (2015). Also, following van der Pas, Szabo, and van der Vaart (2017), we have been able to derive credible sets for multivariate normal means under global-local priors.

188. Fast Parameter Estimation of GEV using Neural Networks

[Student Paper Competition 1, (page 5)]

Sweta RAI, *Colorado School of Mines*

The heavy-tailed behavior of the generalized extreme-value distribution makes it a popular choice for modeling extreme events such as floods, droughts, heatwaves, wildfires, etc. However, estimating the distribution's parameters using conventional maximum likelihood methods can be computationally intensive, even for moderate-sized datasets. To overcome this limitation, we propose a computationally efficient, likelihood-free estimation method utilizing a neural network. Through an extensive simulation study, we demonstrate that the proposed neural network-based method provides Generalized Extreme Value (GEV) distribution parameter estimates with comparable accuracy to the conventional maximum likelihood method but with a significant computational speedup. To account for estimation uncertainty, we utilize parametric bootstrapping, which is inherent in the trained network. Finally, we apply this method to 1000-year annual maximum temperature data from the Community Climate System Model version 3 (CCSM3) across North America for three atmospheric concentrations: 289 ppm CO_2 (pre-industrial), 700 ppm CO_2 (future conditions), and 1400 ppm CO_2, and compare the results with those obtained using the maximum likelihood approach.

189. A Curious Distribution

[03.M1.I31, (page 20)]

Marepalli RAO, *University of Cincinnati*
Zhaochong YU, *University of Cincinnati*
Neelakshi CHATTERJEE, *University of Cincinnati*
NONE,

This problem arose from a paper of Persi Diaconis and Ronald Graham on Guessing the Guessing published in the American Mathematical Monthly, 2023. A set of n cards numbered serially from 1 to n is laid out face down on a table in a line. The task is to guess the number on each card. Several strategies of guessing under different information modalities were discussed by Diaconis and Graham. In this presentation, we discuss one strategy under no information modality. Select one permutation at random from the set of all permutations and lay it on top of the cards. Let Sn denote the number of correct guesses. We observe some strange properties of the distribution of Sn. Some applications to Clinical Trials will be pointed out.

190. Robust and replicable supervised and unsupervised learning methods for cancer precision medicine

[01.A1.I16, (page 10)]

Naim RASHID, *Department of Biostatistics, Gillings School of Global Public Health, UNC-CH*

The replicability of statistical algorithms for clinical decision-making has been of significant concern in biomedical research, where multiple factors may limit the generalizability of models trained on individual studies. In this talk we illustrate the impact of such factors on existing strategies for training prediction models, motivated by recent problems in the cancer genomics. To improve replicability, we discuss several multi-study learning approaches for supervised and unsupervised learning problems. We show that such approaches yield improved prediction accuracy in new studies in the context of supervised learning, and yield more robust subgroups and gene biomarkers in the context of unsupervised learning. Applications to subtype discovery and prediction in pancreatic cancer will be discussed.

191. Graph estimation in high dimensional time-series

[01.A2.I19, (page 12)]
Arkaprava ROY, *University of Florida*
Anindya ROY, *UMBC*
Subhashis GHOSAL, *NCSU*

Multivariate time series data are routinely collected in many application areas. Although stationarity is very useful modeling assumptions for any time series data, methodological developments are limited under these assumptions for multivariate time series. Under some assumptions on the autocovariance matrices, in this article we achieve those properties for a new class of Gaussian multivariate time series. In this proposed class, the normalized multivariate time series is assumed to be some orthogonal rotation of a set of independent univariate latent time series. To capture the graphical dependence structure among the variables we also propose to sparsely estimate the marginal precision matrix and develop related computational methodologies. An efficient Markov Chain Monte Carlo (MCMC) algorithm is developed for posterior computation. We also study theoretical consistency properties. We show excellent performance in simulations and real data applications.

192 . A Powerful Detection Rule in High Dimensional Mediation Hypothesis Testing

[03.M2.C1, (page 22)]
Asmita ROY, *Texas A&M University*
Xianyang ZHANG, *Texas A&M University*

In genome-wide epigenetic studies, exposures (e.g Single Nucleotide Polymorphisms) affect outcomes (e.g disease status) through intermediate variables like gene expression and DNA methylation. Mediation Analysis offers a way to study these intermediate variables and identifying the presence or absence of causal mediation effects. The existing methods (Joint Significance test, Wald type Sobel's test) are underpowered owing to the composite null hypothesis and the multiple testing burden. We introduce a mediation analysis technique called MLFDR (Mediation Analysis using Local FDR) which uses the local FDRs based the coefficients of the structural equation model specifying the mediation relationship to construct a rejection region. We have shown theoretically as well as through simulation studies that in high-dimensional setting, this method of identifying the mediating variables control the FDR asymptotically and perform better with respect to power than many existing methods like DACT (Liu et al) and JS-mixture (Dai et al).

193 . Dynamic enrichment of Bayesian small sample, sequential, multiple assignment randomized trial (snSMART) design using natural history data: A case study from Duchenne muscular dystrophy

[Special Invited Session 4, (page 18)]
Satrajit ROYCHOUDHURY, *Pfizer Inc.*

In Duchenne muscular dystrophy (DMD) and other rare diseases, recruiting patients into clinical trials is challenging. Additionally, assigning patients to long-term, multi-year placebo arms raises ethical and trial retention concerns. This poses a significant challenge to the traditional sequential drug development paradigm. In this article, we propose a small sample, sequential, multiple assignment, randomized trial (snSMART) design that combines dose selection and confirmatory assessment into a single trial. This multi-stage design evaluates the effects of multiple doses of a promising drug and rerandomizes patients to appropriate dose levels based on their stage 1 dose and response. Our proposed approach increases the efficiency of treatment effect estimates by i) enriching the placebo arm with external control data, and ii) using data from all stages. Data from external control and different stages are combined using a robust Meta-Analytic Combined (MAC) approach to consider the various sources of heterogeneity and potential selection bias. We reanalyze data from a DMD trial using the proposed method and external control data from the Duchenne Natural History Study (DNHS). Our method's estimators show improved efficiency compared to the original trial. Also, the robust MAC-snSMART method most often provides more accurate estimators than the traditional analytic method. Overall, the proposed methodology provides a promising candidate for efficient drug development in DMD and other rare diseases.

194 . Discussions and Challenges in Multi-Arm Multi-Stage (MAMS) Designs

[01.A1.I15, (page 9)]
Niladri ROY CHOWDHURY, *Bristol-Myers-Squibb*
Xue WU, *Penn State University*
Arun KUMAR, *Bristol-Myers-Squibb*

Multi-Arm Multi-Stage (MAMS) clinical trials are seamless Phase 2/3 designs that have multiple treatment arms and concurrent control arm. These designs allow treatment selection at early stage and uses these early stage patients' data in making confirmatory decision at the later stage; thus, saving time and resources. While the designs have been around for some time, they are not popular because of two major challenges. The first challenge is controlling the overall Type I error which is more complicated than traditional two-arm design due to various multiplicity issues. The second challenge is the choice of appropriate endpoint for treatment selection and how that choice may impact the confirmatory outcome. In this presentation, we will discuss these challenges and provide insights on the solutions that one may use to overcome these challenges. Simulation studies will be presented to show benefits of the proposed solutions.

195 . Characterizing Asymptotic Dependence between a Satellite Precipitation Product and Station Data in the Northern US Rocky Mountains via the Tail Dependence Regression Framework with a Gibbs Posterior Inference Approach

[01.M1.I4, (page 4)]

Brook RUSSELL, *Clemson University School of Mathematical and Statistical Sciences*
Yiren DING, *Clemson University School of Mathematical and Statistical Sciences*
Whitney HUANG, *Clemson University School of Mathematical and Statistical Sciences*
Jamie DYER, *Department of Geosciences, Mississippi State University*

The use of satellite precipitation products (SPP) allows for precipitation information to be collected nearly globally, but questions remain regarding their ability to reproduce extreme precipitation over mountainous terrain. In this work, we assess the ability of the Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks-Climate Data Record (PERSIANN-CDR) to capture daily precipitation extremes by comparing PERSIANN-CDR versus corresponding station data in the summer at remote locations in the northern US Rocky Mountains of Wyoming, Idaho, and Montana. Our procedure utilizes the regular variation framework from extreme value theory (EVT), and consists of two distinct approaches. We first assess PERSIANN-CDR's ability to model precipitation extremes through inference on an asymptotic dependence parameter, and conclude that the level of asymptotic dependence is moderate throughout the region. After investigating the degree to which elevation and topographic heterogeneity impact the level of asymptotic dependence, we then suggest a novel approach to identify the impact of a set of meteorological covariates that when combined with the PERSIANN-CDR output, yield an increased level of asymptotic dependence with station data. Our procedure utilizes the tail dependence regression modeling framework and a Gibbs posterior approach for inference, and is able to identify several meteorological covariates that may be important predictor variables.

196 . Limit theorems for semi-discrete optimal transport maps

[Student Paper Competition 2, (page 7)]

Ritwik SADHU, *Department of Statistics and Data Science, Cornell.University*

We study statistical inference for the optimal transport (OT) map (also known as the Brenier map) from a known absolutely continuous reference distribution onto an unknown finitely discrete target distribution. We derive limit distributions for the Lp-estimation error with arbitrary p [1, ∞) and for linear functionals of the empirical OT map. The former has a non-Gaussian limit, while the latter attains asymptotic normality. For both cases, we also establish consistency of the nonparametric bootstrap. The derivation of our limit theorems relies on new stability estimates of functionals of the OT map with respect to the dual potential vector, which could be of independent interest.

197 . Model Selection in Data Integration

[03.M2.I32, (page 21)]

Takumi SAEGUSA, *University of Maryland*

We study model selection problems for merged data from multiple overlapping data sources. In public health data integration, studies to be combined have different target populations with overlaps. Also, subjects in a disease registry appear in other clinical studies as patients. A setting we consider is characterized by (1) duplication of the same units in multiple samples, (2) unidentified duplication across samples, (3) dependence due to finite population sam-

pling. Applications include data synthesis of clinical trials, epidemiological studies, disease registries and health surveys. To address these issues, we propose new information criteria based on the weighted likelihood motivated by Hartley's estimator. Because data integration increases the possibility of overfitting compared to the independent and identically distributed sample, the proposed criteria quantify additional randomness in the data-dependent penalty terms. This penalty accounts for heterogeneity and bias in multiple data sets and guarantees generalizability of scientific findings from combined data. Our results are illustrated with simulation studies and a real data example.

198. Evaluation of methods for analyzing the impact of crossover in Oncology trials

[01.E1.I23, (page 14)]

Abhijoy SAHA, *Eli Lilly and Company*

One-way crossover in Oncology trials usually refers to a situation where patients in the control group of a randomized clinical trial can switch to the experimental treatment arm after a certain predefined criteria has been met per protocol. For example: after disease progression; upon successful interim analysis for patients without disease progression on control arm to receive the experimental beneficial treatment; etc. This impacts estimation of overall survival and introduces bias due to confounding. There are multiple statistical techniques that aim to account for treatment-switching and adjust for the confounding effect in such situations where time-to-event data is being analyzed. We conduct extensive simulation studies by varying different study design parameters to evaluate these techniques and assess their strengths, limitations and appropriateness under various scenarios. We also model the correlation between time-to-event endpoints in our simulation settings to accurately replicate real-world scenarios.

199. Overcoming computational complexities of large-scale spatial modeling with Nearest Neighbor Gaussian Processes using the BRISC R-package

[03.A1.I37, (page 24)]

Arkajyoti SAHA, *University of Washington, Department of Statistics*

Abhirup DATTA, *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health*

In the modern age of big data, much emphasis has been put on managing the computational complexities of spatiotemporal modeling algorithms in Environmental Sciences. Though Gaussian Process has been a go-to choice among practitioners due to its flexibility in modeling location-and-time-indexed datasets, its cubic computational complexity makes it infeasible for large datasets. Among a plethora of approaches that propose a scalable approximation of GP, Nearest neighbor Gaussian Process (NNGP) has established itself as one of the front runners with its multifaceted applications. NNGP models the correlation using information from nearest neighbors, which results in a sparse approximation of the Cholesky factor of the precision matrix with linear computational complexity. We present the R package BRISC, which has been tailored to harness the power of this linear order computation in an embarrassingly parallel framework. BRISC exploits the sparsity structure for estimation, prediction, bootstrap-based inference of spatial parameters, simulation of large spatial data from GP, and prediction in spatial probit models. BRISC is also utilized in spatial machine learning for non-linear estimation of means in GP with Random Forests.

200. A Bayesian semiparametric model for variable selection in compositional data

[03.E1.I47, (page 28)]

Satabdi SAHA, *The University of Texas MD Anderson Cancer Center*

Christine PETERSON, *The University of Texas MD Anderson Cancer Center*

Abstract: Ongoing advances in metagenomic sequencing have allowed unprecedented insights into the molecular activities of microbial communities. This has fueled a strong scientific interest in understanding the critical role the microbiome plays in governing human health, by identifying microbial features associated to clinical outcomes of interest. Several features of the microbiome data, however, limit the ability of existing variable selection approaches. High rates of zero-inflation, compositionality and existence of phylogenetic associations among the bacterial taxa pose major challenges. To address the above challenges, we propose novel priors that simultaneously encourages aggregation and selection of relevant microbiome features while accounting for data compositionality. We demonstrate that our proposed method outperforms existing penalized approaches

for microbiome variable selection in both simulation and the real data studies.

201 . Bayesian Group Sparsity and Smoothing on Graphs

[04.M1.I51, (page 31)]

Huiyan SANG, *Texas A&M University*

Graphs have been commonly used to represent complex data structures. In models dealing with graph-structured data, multivariate parameters may not only exhibit sparse patterns but have structured sparsity and smoothness in the sense that both zero and non-zero parameters tend to cluster together. We propose a new prior for high-dimensional parameters with graphical relations, referred to as the Tree-based Low-rank Horseshoe (T-LoHo) model, that generalizes the popular univariate Bayesian horseshoe shrinkage prior to the multivariate setting to detect structured sparsity and smoothness simultaneously. The T-LoHo prior can be embedded in many high-dimensional hierarchical models. To illustrate its utility, we apply it to regularize a Bayesian high-dimensional regression problem where the regression coefficients are linked by a graph, so that the resulting clusters have flexible shapes and satisfy the cluster contiguity constraint with respect to the graph. We design an efficient Markov chain Monte Carlo algorithm that delivers full Bayesian inference with uncertainty measures for model parameters such as the number of clusters. We offer theoretical investigations of the clustering effects and posterior concentration results. Finally, we illustrate the performance of the model with simulation studies and a real data application for anomaly detection on a road network. The results indicate substantial improvements over other competing methods such as the sparse fused lasso.

202 . Bayesian Semiparametric Covariate Informed Multivariate Density Deconvolution

[04.M2.I60, (page 34)]

Abhra SARKAR, *The University of Texas at Austin*

Estimating the marginal and joint densities of the long-term average intakes of different dietary components is an important problem in nutritional epidemiology. Since these variables cannot be directly measured, data are usually collected in the form of 24-hour recalls of the intakes. The problem of estimating the density of the latent long-term average intakes from their observed but error-contaminated recalls then becomes a problem of multivariate deconvolution of densities. The underlying densities could potentially vary with the subjects' demographic characteristics such as sex, ethnicity, age, etc. The problem of density deconvolution in the presence of associated precisely measured covariates has, however, never been considered before, not even in the univariate setting. We present a flexible Bayesian semiparametric approach to covariate-informed multivariate deconvolution. Building on recent advances in copula deconvolution and conditional tensor factorization techniques, our proposed method not only allows the joint and the marginal densities to vary flexibly with the associated predictors but also allows automatic selection of the most influential predictors. Importantly, the method also allows the density of interest and the density of the measurement errors to vary with potentially different sets of predictors. We design Markov chain Monte Carlo algorithms that enable efficient posterior inference, appropriately accommodating uncertainty in all aspects of our analysis. The empirical efficacy of the proposed method is illustrated through simulation experiments. Its practical utility is demonstrated in the afore-described nutritional epidemiology applications in estimating covariate-adjusted long-term intakes of different dietary components. An important by-product of the approach is a solution to covariate-informed ordinary multivariate density estimation.

203 . Bootstrapping the Error of Oja's Algorithm

[Special Invited Session 6, (page 23)]

Purnamrita SARKAR, *University of Texas at Austin*

We consider the problem of quantifying uncertainty for the estimation error of the leading eigenvector from Oja's algorithm for streaming principal component analysis, where the data are generated IID from some unknown distribution. By combining classical tools from the U-statistics literature with recent results on high-dimensional central limit theorems for quadratic forms of random vectors and concentration of matrix products, we establish a weighted approximation result for the error between the population eigenvector and the output of Oja's algorithm. Since estimating the covariance matrix associated with the approximating distribution requires knowledge of unknown model parameters, we propose a multiplier bootstrap algorithm that may be updated in an online manner. We establish conditions under which the

bootstrap distribution is close to the corresponding sampling distribution with high probability, thereby establishing the bootstrap as a consistent inferential method in an appropriate asymptotic regime.

204. Joint Point and Variance Estimation under a Hierarchical Bayesian model for Survey Count Data

[01.A2.I18, (page 11)]

Terrance SAVITSKY, *BLS*

TBA

205. Estimation of finite population proportions for small areas – a statistical data integration approach

[Poster Session, (page 17)]

Aditi SEN, *University of Maryland*

Empirical best prediction (EBP) is a well-known method for producing reliable proportion estimates when the primary data source provides only small or no sample from finite populations. There are at least two potential challenges encountered in implementing the existing EBP methodology. First, one must accurately link the sample to the finite population frame. This may be a difficult or even impossible task because of absence of identifiers that can be used to link sample and the frame. Secondly, the finite population frame typically contains limited auxiliary variables, which may not be adequate for building a reasonable working predictive model. We propose a data linkage approach in which we replace the finite population frame by a big sample that does not have the outcome binary variable of interest, but has a large set of auxiliary variables. Our proposed method calls for fitting the assumed model using data from the smaller sample, imputing the outcome variable for all the units of the big sample, and then finally using these imputed values to obtain standard weighted proportion for the finite population. We develop a new adjusted maximum likelihood method to avoid estimates of model variance on the boundary encountered in the commonly used in maximum likelihood estimation method. We propose a consistent estimator of mean squared prediction error (MSPE) using a parametric bootstrap method and address computational issues by developing efficient EM algorithm. We illustrate the proposed methodology in the context of election projection for small areas.

206. Empirical partially Bayes multiple testing and compound Chi-square decisions

[01.E1.I25, (page 14)]

Bodhisattva SEN, *Columbia University*
Nikolaos IGNATIADIS, *Columbia University*

We study multiple testing in the normal means problem with estimated variances that are shrunk through empirical Bayes methods. The situation is asymmetric in that a prior is posited for the nuisance parameters (variances) but not the primary parameters (means). If the prior were known, one could proceed by computing p-values conditional on sample variances; a strategy called partially Bayes inference by Sir David Cox. These conditional p-values satisfy a Tweedie-type formula and are approximated at nearly-parametric rates when the prior is estimated by nonparametric maximum likelihood. If the variances are in fact fixed, the approach retains type-I error guarantees.

207. Incorporating network side information into supervised learning

[03.A1.I38, (page 25)]

Subhabrata SEN, *Harvard University*
Sagnik NANDY, *University of Pennsylvania*
Subhabrata SEN, *Harvard University*

In this talk, we will discuss supervised learning with network side information, where the network represents additional side information regarding the underlying model parameters. This setting is especially common in biological applications, where the network captures independent domain knowledge. Incorporation of this side information is expected to improve model estimation accuracy and aid in the discovery of non-null variables. We will formulate this question using a simple Bayesian generative model, and introduce an algorithm for Bayes optimal recovery in this setting. We will demonstrate the superior performance of our algorithm compared to existing benchmarks using numerical experiments. Based on joint work with Sagnik Nandy (UPenn).

208. Anomalous clique detection and identification in inhomogeneous networks

[03.M1.I30, (page 20)]

Srijan SENGUPTA, *North Carolina State University*
Subhankar BHADRA, *North Carolina State University*

Cliques or fully connected subgraphs are the most well-studied local structures in the network science literature. We consider the following problem: given a large background network, we want to know whether the network contains an anomalous clique (the detection problem). If the answer to this question is yes, then we further want to identify the members of this anomalous clique (the identification problem). We propose an inferential tool based on egonets to answer both questions. The proposed method is computationally efficient, naturally amenable to parallel computing, and easily extends to a wide variety of network models. We derive theoretical guarantees for both the detection and identification tasks under a variety of inhomogeneous random graph models. We demonstrate through simulation studies that the egonet method works well under a wide variety of network models. We obtain some interesting empirical results by applying the egonet method to several well-studied benchmark datasets.

209 . Connections between stick-breaking measures and Markov chains

[Memorial Session 3, (page 28)]

Sunder SETHURAMAN, *University of Arizona*

In this talk, we introduce a notion of 'Markovian stick-breaking' measures, which is a generalization of the Dirichlet process, and discuss connections between these measures and (i) time-inhomogeneous Markov chains, related to the Metropolis algorithm, as well as (ii) multi-state promoter models of mRNA dynamics.

The talk will be based on the papers with Zach Dietz, William Lippitt, and Xueying Tang.

210. Spatial scale-aware tail dependence modeling for high-dimensional spatial extremes

[04.M1.I56, (page 33)]

Ben SHABY, *Colorado State University*

TBA

211 . Predicting Rebel Movements in Mexico: A Machine Learning Approach

[Poster Session, (page 17)]

Harsh Hemant SHAH, *University of Minnesota*

Rebels have caused anti-government protests and general anarchy. Researchers are developing a new methodology that uses machine learning to anticipate anti-government movements and identify significant predictors of conflicts. The current analysis of conflicts for a specific demographic location with spatial and temporal relationships needs to be improved. The use of machine learning and deep learning has successfully anticipated election disputes, but it lacks spatial structure and cannot explain relevant predictors. The researchers are investigating the spatiotemporal relationships and key determinants that support rebel movements in Mexico.

212 . Statistical Considerations for Externally Controlled Studies using Real World Data with Small Sample Size

[03.M1.I29, (page 20)]

Gaurav SHARMA, *Takeda Pharmaceuticals*
Jo GAO, *Takeda Pharmaceuticals*

Use of Real World Data (RWD) sources (e.g. EHRs, claims data, registries) can reduce the patient burden and possibly accelerate the drug development process. This is especially true for rare diseases with unmet medical needs, where due to ethical or feasibility reasons, RWD are sometimes used as external control arms to compare with data from prospectively treated single-arm trials. Many statistical challenges arise when utilizing RWD for external controls which have been part of the FDA guidance documents over the years. In this presentation, we will discuss specifically two key challenges with utilizing external controls are unbalanced samples and unequal patient encounters between the external control arm vs single arm trials. In such a setting, interval censored data with different assessment time structures between the two arms are produced, making right censoring inappropriate while cohort balancing is necessary. Cohort balancing approaches along with the stratified generalized log-rank test for interval censored data will be discussed. Numerical simulation results comparing type I error, power, and other diagnostic metrics for different methods will be presented in the context of small to moderate sample sizes.

213 . Probabilistic Inverse Model: An Application in Hydrology

[Student Paper Competition 1, (page 5)]

Somya SHARMA, *University of Minnesota - Twin Cities*

Rapid advancement in inverse (machine learning) modeling methods have brought into light their sus-

ceptibility to imperfect data. This has made it imperative to obtain more explainable and trustworthy estimates from these models. In hydrology, basin characteristics can be noisy or missing, impacting streamflow prediction. We propose a probabilistic inverse model framework that can reconstruct robust hydrology basin characteristics from dynamic input weather driver and streamflow response data. We address two aspects of building more explainable inverse models, uncertainty estimation (uncertainty due to imperfect data and imperfect model) and robustness. This can help improve the trust of water managers, handling of noisy data and reduce costs. We also propose an uncertainty-based loss regularization that offers removal of 17% of temporal artifacts in reconstructions, 36% reduction in uncertainty and 4% higher coverage rate for basin characteristics. The forward model performance (streamflow estimation) is also improved by 6% using these uncertainty learning based reconstructions.

214 . Sequential change detection via backward confidence sequences

[03.A2.I41, (page 26)]

Shubhanshu SHEKHAR, *Carnegie Mellon University*
Aaditya RAMDAS, *Department of Statistics and Data Science, Carnegie Mellon University*

We present a simple reduction from sequential estimation to sequential changepoint detection (SCD). In short, suppose we are interested in detecting changepoints in some parameter or functional θ of the underlying distribution. We demonstrate that if we can construct a confidence sequence (CS) for θ, then we can also successfully perform SCD for θ. This is accomplished by checking if two CSs — one forwards and the other backwards — ever fail to intersect. Since the literature on CSs has been rapidly evolving recently, the reduction provided in this paper immediately solves several old and new change detection problems. Further, our "backward CS", constructed by reversing time, is new and potentially of independent interest. We provide strong nonasymptotic guarantees on the frequency of false alarms and detection delay, and demonstrate numerical effectiveness on several problems.

215. On the Realization Problem of Tail Dependence Matrices

[01.E1.I24, (page 14)]

Narikadu D. SHYAMALKUMAR, *University of Iowa*

The tail dependence coefficient is a popular bivariate tail dependence measure. Akin to the correlation matrix, a multivariate tail dependence measure is the tail dependence matrix (TDM) constructed using these bivariate measures. The problem of determining whether a given d $\times$ d matrix is a TDM, the realization problem, is significantly more complex than determining it is a correlation matrix. Using an LP formulation, we show that the combinatorial structure of the constraints is related to the intractable max-cut problem in a weighted graph. This connection provides an avenue for constructing parametric classes admitting a polynomial in d algorithm for determining membership in its constraint polytope. We show how the inherent symmetry and sparsity in the parametrization of a class of TDMs help significantly simplify the LP formulation, leading to polynomial complexity of its constrained realization problem - with some O(1) in complexity. We also study the subset of TDMs supported by the t-copula family. This is joint work with Siyang Tao (Ball State University).

216. Misspecification and Calibration effects in Sequential decision making

[Plenary Lecture 3, (page 17)]

Aarti SINGH, *Carnegie Mellon University*

The most widespread use of statistics and machine learning comes from the ability to map complex input-output associations based on some given data. In real-world deployments, these mappings are typically not done once with fixed data. Rather, data-driven systems need to continually interact with their environment, not only to make sequential predictions, but often also make decisions about what data to collect to improve their performance. This talk will focus on bandit optimization algorithms, a key approach for sequential decision making, and discuss two real-world challenges - misspecification of the underlying data generating function and calibration effects where repeated querying changes the function being optimized.

217 . Anytime-Valid Confidence Sequences in an Enterprise A/B Testing Platform

[03.A2.I41, (page 26)]

Ritwik SINHA, *Adobe Research*

A/B tests are the gold standard for evaluating digital experiences on the web. However, traditional

"fixed-horizon" statistical methods are often incompatible with the needs of modern industry practitioners as they do not permit continuous monitoring of experiments. Frequent evaluation of fixed-horizon tests ("peeking") leads to inflated type-I error and can result in erroneous conclusions. We have released an experimentation service on the Adobe Experience Platform based on anytime-valid confidence sequences, allowing for continuous monitoring of the A/B test and data-dependent stopping. We demonstrate how we adapted and deployed asymptotic confidence sequences in a full featured A/B testing platform, describe how sample size calculations can be performed, and how alternate test statistics like "lift" can be analyzed. On both simulated data and thousands of real experiments, we show the desirable properties of using anytime-valid methods instead of traditional approaches.

218. How Advanced Statistical and Data Science Methods Are Reshaping Next-generation Psychometrics

[03.M1.I28, (page 19)]

Sandip SINHARAY, *Educational Testing Service*

If Statistics is a big river, then "Psychometrics" is the small river that comprises the collection of quantitative methods and models that are developed in psychological and educational testing and goes down the stream to the big river of Statistics. As "next-generation" and computerized assessment systems generate complex data that record the individuals' interactions with the tasks or items at finer time granularity than in the past, the volume, velocity, and variety of the data pose new challenges to researchers in psychometrics to analyze them in order to make useful conclusions. In this presentation, I will discuss how current psychometric methodology is being extended, with the help of advanced statistics and data science methods, to handle the data originating from the next-generation assessment systems. I will also discuss how current educational tests are being criticized for perpetuating social injustice through their content, design, and use, and discuss how an ongoing NSF-funded project is aiming to partially address that criticism.

219. The edge of discovery: Controlling the local false discovery rate at the margin

[01.E1.I25, (page 14)]

Jake SOLOFF, *University of Chicago*
Daniel XIANG, *University of Chicago*
William FITHIAN, *University of California, Berkeley*

Despite the popularity of the false discovery rate (FDR) as an error control metric for large-scale multiple testing, its close Bayesian counterpart the local false discovery rate (lfdr), defined as the posterior probability that a particular null hypothesis is false, is a more directly relevant standard for justifying and interpreting individual rejections. However, the lfdr is difficult to work with in small samples, as the prior distribution is typically unknown. We propose a simple multiple testing procedure and prove that it controls the expectation of the maximum lfdr across all rejections; equivalently, it controls the probability that the rejection with the largest p-value is a false discovery. Our method operates without knowledge of the prior, assuming only that the p-value density is uniform under the null and decreasing under the alternative. We also show that our method asymptotically implements the oracle Bayes procedure for a weighted classification risk, optimally trading off between false positives and false negatives. We derive the limiting distribution of the attained maximum lfdr over the rejections, and the limiting empirical Bayes regret relative to the oracle procedure.

220. Statistics and the Fair Administration of Justice: Assessing Bloodstain Pattern Evidence

[Special Invited Session 6, (page 23)]

Hal STERN, *University of California Irvine*
Tong ZOU, *University of California Irvine*

Statistics has emerged as a critical topic in ongoing discussions regarding the use of science to assess forensic evidence. A 2009 U.S. National Academies report on forensic science and a subsequent 2016 report by the U.S. President's Council of Advisers on Science and Technology raised questions about the scientific underpinnings for the analysis of a number of types of forensic evidence. Misapplication of forensic science has been identified as a contributing factor in nearly half of 362 cases in which DNA helped exonerate wrongly-convicted individuals. This talk provides some background on forensic statistics and demonstrates approaches to inference for bloodstain pattern evidence. Contributions include a novel approach to representing the bloodstain patterns and the application of a Dirichlet Process Mixture Model for assessing the likelihood of observing a given pat-

tern under different causal mechanisms.

221. Model selection for network data based on spectral information

[01.M2.I9, (page 6)]

Jonathan STEWART, *Florida State University*
Jairo PEÑA, *Florida State University*

The statistical network analysis literature offers a plethora of models for network data, but lacks comprehensive methodology for model selection. Much of what has been proposed in the way of model selection in the literature has focused on specific classes of models, e.g., selecting the number of clusters in a stochastic block model or relevant model terms from an exponential-family random graph model. One of the foremost challenges lies in developing methods for model selection that allow for candidate models to come from different model classes and frameworks. We present a novel methodology for model selection in the context of modeling network network data which exploits the information contained in the spectrum of the Laplacian matrix. Our method is purely non-parametric and model-free, in the sense that we require only the ability to reliably and efficiently simulate networks from models. The performance of our proposed methodology is explored for popular classes of network data models in simulation, where we demonstrate empirical evidence that our methods are able to consistently select data-generating models in important classes of models. We apply our methods to two applications, one to a collaboration network of network science researchers and to the Sampson monk network.

222. Predicting the anti-government conflicts based on Spatio-Temporal Geo-Political data from Mexico

[Poster Session, (page 17)]

Vishal SUBEDI, *University of Minnesota*

Rebels have caused chaos and anti-government movements. There have been significant developments in predicting conflicts, but not when having a spatial as well as temporal correlation with its neighboring places and previous conflicts respectively. Machine learning, including deep learning, has been effective in predicting electoral conflicts using word embedding but lacks spatial structure and explanatory power. We are developing a novel methodology to predict anti-government rebel movements in Mexico ahead of time. This will help policymakers make informed decisions and improve the distribution of armed personnel. Initial results from black box models and Random Forest are significant. The study is also exploring spatio-temporal relationships and important predictors.

223. RESPONSE MODEL SELECTION IN CASE OF NOT MISSING AT RANDOM NONRESPONSE WITH APPLICATION TO REAL DATA

[03.A2.I44, (page 27)]

Michael SVERCHKOV, *Bureau of Labor Statistics*

Sverchkov and Pfeffermann (S-P, 2008, 2018) consider estimation under informative sampling and not missing at random (NMAR) nonresponse. To account for the nonresponse, S-P assume a given response model and estimate the corresponding response probabilities by application of the Missing Information Principle, which consists of defining the likelihood as if there was complete response and then integrating out the unobserved outcomes from the likelihood employing the relationship between the sample and sample-complement distributions. A key condition for the success of this approach is the specification of the response model. In this presentation we consider information criteria based on the above likelihood, and show how they can be used for the selection of the response model. We illustrate the approach on a real data example.

Key words: information criteria, likelihood ratio tests, missing information principle, nonresponse bias

REFERENCES Sverchkov, M. (2008). A new approach to estimation of response probabilities when missing data are not missing at random. Joint Statistical Meetings, Proceedings of the Section on Survey Research Methods, 867-874. Sverchkov, M. and Pfeffermann, D. (2018). Small area estimation under informative sampling and not missing at random non-response. Journal of Royal Statistical Society, ser. A, 181, Part 4, pp. 981–1008.

224. AI/ML in retail digital transformation

[03.A2.I43, (page 27)]

Shirish TATIKONDA, *Walmart Data Ventures*

TBA

225. Ancestral Inference for Bellman Harris process.

[Memorial Session 2, (page 26)]
Anand VIDYASHANKAR, *George Mason University*

In this presentation, I will give a brief description of Prof. Athreya's contributions to probability, statistics, and applications. I will then explain in some detail how some of the tools he developed can be used when studying the ancestral inference problem.

226. Leveraging Real-World Data and Real-World Evidence in Clinical Trial Design and Analysis and its Causal Implications

[03.M1.I29, (page 20)]
Chenguang WANG, *Senior Director, Regeneron Pharmaceuticals*

Incorporating real-world data (RWD) in regulatory decision-making demands much more than "mixing" RWD with investigational clinical trial data. The RWD must undergo appropriate analysis for deriving the right real-world evidence (RWE). Moreover, such analysis should be integrated with the design and analysis of the investigational study for regulatory decision-making. The standard clinical trial toolbox does not offer ready solutions for such tasks. Therefore, there is an unmet need for sound clinical trial design and analysis for leveraging RWE in clinical evaluations in the context of regulatory decision-making.

Recently, we proposed a method that extended the Bayesian power prior approach for leveraging external RWD. The method used propensity score to pre-select a subset of RWD patients that were "similar" to those in the current study, and to stratify the selected patients together with those in the current study into more homogeneous strata. In this talk, we will review this propensity score-integrated power prior approach and discuss the underlying causal assumptions the method requires.

227. Coverage of Credible Intervals in Bayesian Multivariate Isotonic Regression

[Student Paper Competition 2, (page 8)]
Kang WANG, *North Carolina State University*

We consider the nonparametric multivariate isotonic regression problem, where the regression function is assumed to be nondecreasing with respect to each predictor. Our goal is to construct a Bayesian credible interval for the function value at a given interior point with assured limiting frequentist coverage. We put a prior on unrestricted step-functions, but make inference using the induced posterior measure by an "immersion map" from the space of unrestricted functions to that of multivariate monotone functions. This allows maintaining the natural conjugacy for posterior sampling. A natural immersion map to use is a projection via a distance, but in the present context, a block isotonization map is found to be more useful. The approach of using the induced "immersion posterior" measure instead of the original posterior to make inference provides a useful extension of the Bayesian paradigm, particularly helpful when the model space is restricted by some complex relations. We establish a key weak convergence result for the posterior distribution of the function at a point in terms of some functional of a multi-indexed Gaussian process that leads to an expression for the limiting coverage of the Bayesian credible interval. Analogous to a recent result for univariate monotone functions, we find that the limiting coverage is slightly higher than the credibility, the opposite of a phenomenon observed in smoothing problems. Interestingly, the relation between credibility and limiting coverage does not involve any unknown parameter. Hence by a recalibration procedure, we can get a predetermined asymptotic coverage by choosing a suitable credibility level smaller than the targeted coverage, and thus also shorten the credible intervals.

228. Opportunities and challenges in neuroscience clinical trials

[01.M2.I7, (page 6)]
Ling WANG, *Alkermes Inc*

Neuroscience drug development is a very dynamic area that includes many disease domains, from common conditions such as alzheimer's, depression to rare diseases including ALS. We will discuss recent development in new therapies in neuroscience and statistical challenges and opportunities in this area.

229. Double-robust Bayesian variable selection and model prediction with spherically symmetric errors

[04.M2.I57, (page 33)]
Min WANG, *University of Texas-San Antonio*

TBA

230 . Multiply-robust estimation of causal treatment effect on a binary outcome with integrated information from secondary outcomes

[03.M1.I31, (page 21)]

Ming WANG, *Case Western Reserve University*
Chixiang CHEN, *University of Maryland*
Shuo CHEN, *University of Maryland*
Qi LONG, *University of Pennsylvania*
THERE IS ANOTHER CO-AUTHOR: SUDESHNA DAS FROM MASSACHUSETTS GENERAL HOSPITAL, HARVARD MEDICAL SCHOOL,

An assessment of the causal treatment effect in the development and progression of certain diseases is important in clinical trials and biomedical studies. However, it is not possible to infer a causal relationship when the treatment assignment is imbalanced and confounded by other mechanisms. Specifically, when the treatment assignment is not randomized and the primary outcome is binary, a conventional logistic regression may not be valid to elucidate any causal inference. Moreover, exclusively capturing all confounders is extremely difficult and even impossible in large-scale observational studies. We propose a multiply-robust (MultiR) estimator for estimating the causal effect with a binary outcome, where multiple propensity score models and conditional mean imputation models are used to ensure estimation robustness. Further, we propose an enhanced MultiR (eMultiR) estimator that reduces the estimation variability of MultiR estimates by incorporating secondary outcomes that are highly correlated with the primary binary outcome. The resulting estimates are less sensitive to model mis-specification compared to those based on state-of-the-art methods (e.g., doubly-robust estimators). These estimates are verified through both theoretical and numerical assessments. The utility of (e)MultiR estimation is illustrated using the Uniform Data Set (UDS) from the National Alzheimer's Coordinating Center with the objective to detect the causal effect of the short-term use of antihypertensive medications on the development of dementia or mild cognitive impairment.

231. Multivariate functional data clustering using adaptive density peak detection

[03.M1.I27, (page 19)]

Xiaofeng WANG, *Cleveland Clinic*

Clustering for multivariate functional data is a challenging problem since the data are represented by a set of curves and functions belonging to an infinite-dimensional space. In this article, we propose a novel clustering method for multivariate functional data using an adaptive density peak detection technique. It is a quick cluster center identification algorithm based on the two measures of each functional data observation: the functional density estimate and the distance to the closest observation with a higher functional density. We suggest two types of functional density estimators for multivariate functional data. The first one is a functional k-nearest neighbor density estimator based on (a) an L2 distance between raw functional curves, or (b) a semimetric of multivariate functional principal components. The second one is a k-nearest neighbor density estimator based on multivariate functional principal scores. Our clustering method is computationally fast since it does not need an iterative process. The flexibility and advantages of the method are examined by comparing it with other existing clustering methods in simulation studies. A user-friendly R package FADPclust is developed for public use. Finally, our method is applied to a real case study in lung cancer research.

232. Fire-Atmosphere Coupling Implications for Wildland Fire Modeling and Decision Making

[03.E1.I49, (page 29)]

Joseph WERNE, *NorthWest Research Associates*
Joseph WERNE, *NorthWest Research Associates*
Wayne SPENCER, *Conservation Biology Institute*

Wildfires are becoming increasingly destructive in the western US, and informed management action is needed to minimize their impacts. At the same time, the development of fast and accurate tools to guide decision making is challenging, because nonlinear wildfire behavior is complex, and the net impact of relevant physical effects can sometimes be subtle and counter-intuitive. Fire-atmosphere coupling is a particular area where model improvements are both possible and promising. For example, while forest thinning by removing some trees reduces the forest fuel load, it nevertheless has the potential to increase fire intensity and spread as a result of the concomitant increase in available oxygen as the local forest-stand wind resistance is also reduced. Current operational fire models to evaluate fuels-reduction treatments do not automatically address this balance with

ventilation increase. Similarly, fire-induced buoyancy effects that have the potential to feedback and give rise to so-called "plume fires" and associated potential "megafire" behavior are also absent, because fire-atmosphere coupling is costly to model accurately, and the models that include it are typically too slow to be used in operational settings. While incorporation of archived statistical data associated with past wildfires into operational models may help to include some of these effects, direct application is complicated by shifting climate conditions that may not be reflected in the historical data. Better models are needed to characterize how local atmospheric motions (i.e., the wind field) are affected by vegetation structure, terrain, and other factors, as well as address how fire-induced buoyancy and associated winds affect fire behavior. In this talk I will describe current work we and others are doing to address how forest structure (i.e., the distribution of above-ground biomass) affects air flow. I will discuss how we are using available forest-structure data to model vegetation wind-drag across a forested landscape, how fluid dynamics can be used to estimate windspeed increases associated with vegetation reductions, and how the magnitude of fire-induced winds can similarly be estimated, helping us identify where buoyancy-driven air-mass circulations created by fire may compete with or even exceed the ambient wind field, signaling the need for caution when interpreting current operational-model predictions.

233. Inference on contrasts of monotone functions

[03.E1.I48, (page 29)]

Ted WESTLING, *Department of Mathematics and Statistics, University of Massachusetts Amherst*
Yujian WU, *University of Massachusetts Amherst*
Eric MORENZ, *University of Washington*
Marco CARONE, *University of Washington*

In many settings, it is known that a function of interest is monotone due to probabilistic constraints or prior scientific knowledge. General approaches to estimation of a monotone function have been proposed, the properties of these estimators have been derived, and the results have been applied to a variety of specific problems. In some cases, the contrast of two or more monotone functions is of scientific interest in addition to the functions themselves. For instance, in the context of current status censoring, investigators may be interested in the difference or ratio between the survival functions of the event of interest under two different conditions, such as treatment and control. A natural approach to estimation in such problems is to apply the contrast to monotone estimators. To obtain asymptotically valid inference based on the resulting contrast estimator, the joint asymptotic behavior of the monotone estimators is needed. We derive conditions under which two or more generalized Grenander-type estimators converge jointly in distribution, upon proper centering and scaling, to a multivariate analogue of the Chernoff distribution. We use our general results to obtain large-sample inference for contrasts of monotone functions in several examples with important real-world applications.

234. Multilayered Network Models for Security: Enhancing System Security Engineering with Orchestration

[01.M2.I11, (page 7)]

Adam D. WILLIAMS, *Principal R&D Systems Engineer, Sandia Laboratories*

TBA

235. Fast Generalized Functional Principal Components Analysis

[03.M2.I33, (page 22)]

Julia WROBEL, *Department of Biostatistics and Informatics, Colorado School of Public Health*

TBA

236. Empirical Bayes estimation: When does g-modeling beat f-modeling in theory (and in practice)?

[01.E1.I25, (page 14)]

Yihong WU, *Yale University*
Yandi SHEN, *UChicago*

Empirical Bayes (EB) is a popular framework for large-scale inference that aims to find data-driven estimators to compete with the Bayesian oracle that knows the true prior. Two principled approaches to EB estimation have emerged over the years: f-modeling, which constructs an approximate Bayes rule by estimating the marginal distribution of the data, and g-modeling, which estimates the prior from data and then applies the learned Bayes rule. For the Poisson model, the prototypical examples are the celebrated Robbins estimator and the nonparametric MLE (NPMLE), respectively. It has long been recognized in practice that the Robbins estimator, while being conceptually appealing and computationally

simple, lacks robustness and can be easily derailed by "outliers" (data points that were rarely observed before). In this talk we provide a theoretical justification for the superiority of NPMLE over Robbins for heavy-tailed data by considering priors with bounded pth moment previously studied for the Gaussian model. For the Poisson model with sample size n, assuming $p > 1$ (for otherwise triviality arises), we show that the NPMLE with appropriate regularization and truncation achieves a total regret $\tilde{\Theta}(n^{3/(2p+1)})$, which is minimax optimal within logarithmic factors. In contrast, the total regret of Robbins estimator (with similar truncation) is $\tilde{\Theta}(n^{3/(p+2)})$ and hence suboptimal by a polynomial factor. Joint work with Yandi Shen: https://arxiv.org/abs/2211.12692

237. Statistical Optimality of Federated Learning Beyond Stationary Points

[01.M1.I1, (page 3)]

Jiaming XU, *Duke University*
Lili SU, *Northeastern University*
Pengkun YANG, *Tsinghua University*

Federated Learning (FL) is a promising decentralized learning framework and has great potentials in privacy preservation and in lowering the computation load at the cloud. Recent work showed that FedAvg and FedProx – the two widely-adopted FL algorithms – fail to reach the stationary points of the global optimization objective even for homogeneous linear regression problems. Further, it is concerned that the common model learned might not generalize well locally at all in the presence of heterogeneity.

In this paper, we analyze the convergence and statistical efficiency of FedAvg and FedProx, addressing the above two concerns. Our analysis is based on the standard non-parametric regression in a reproducing kernel Hilbert space (RKHS), and allows for heterogeneous local data distributions and unbalanced local datasets. We prove that the estimation errors, measured in either the empirical norm or the RKHS norm, decay with a rate of $1/t$ in general and exponentially for finite-rank kernels. In certain heterogeneous settings, these upper bounds also imply that both FedAvg and FedProx achieve the optimal error rate.

Based on joint work with Lili Su (Northeastern University) and Pengkun Yang (Tsinghua University). Preprint available at https://arxiv.org/pdf/2106.15216.pdf.

238 . Cross Validation Importance Learning (CVIL)

[04.M1.I54, (page 32)]

Chenglong YE, *University of Kentucky*
Yuhong YANG, *University of Minnesota*

As applied in examples of self-driving cars, virtual personal assistants, and online customer support, machine learning is prevalent. Along with its success and popularity, interpretability is one obstacle that prevents people from fully interpreting and trusting machines' decisions. To help demystify a seemingly uninterpretable algorithm/procedure from a variable importance perspective, we propose two types of variable importance measures (named CVIL) based on cross-validation. For any modeling procedure, CVIL measures the relative difference in the predictive performances after deleting a variable in the data set or replacing the variable with a constant, thus enhancing a variable-level interpretation of any modeling procedure. Together, the two proposed variable importance measures can help discover model structures and causal relationships. Theoretical properties are also shown: 1) Under some mild conditions, CVIL is consistent in the sense that it converges to the proposed theoretical variable importance as the sample size grows. 2) Confidence intervals are also constructed for statistical inference so that follow-up variable selection can be performed in a convenient way (cutoff is easy to determine). The performances of CVIL are demonstrated through a variety of simulation settings and real data examples.

239. Evaluation of Log-rank, RMST and MaxCombo in Immuno-Oncology(IO) trials – A Retrospective Analysis in Patients Treated with Anti-PD1/PD-L1 Agents across Solid Tumors

[01.M2.I8, (page 6)]

Jiabu YE, *Merck Research Labs*

TBA

240. Bayesian modeling with derivative Gaussian processes of event-related potentials

[03.A1.I36, (page 24)]

Cheng-Han YU, *Marquette University*
Meng LI, *Rice University*
Marina VANNUCCI, *Rice University*

We propose a semiparametric Bayesian model to

infer the locations of stationary points of a nonparametric function, which also produces an estimate of the function. We use Gaussian processes as a flexible prior for the underlying function and impose derivative constraints to control the function's shape via conditioning. We develop an inferential strategy that intentionally restricts estimation to the case of at least one stationary point, bypassing possible mis-specifications in the number of stationary points and avoiding the varying dimension problem that often brings in computational complexity. We illustrate the proposed methods using simulations and then apply the method to the estimation of event-related potentials derived from electroencephalography (EEG) signals.We show how the proposed method automatically identifies characteristic components and their latencies at the individual level, which avoids the excessive averaging across subjects that is routinely done in the field to obtain smooth curves.

241. Balancing Weights for Causal Inference in Observational Factorial Studies

[01.M1.I2, (page 3)]

Ruoqi YU, *University of California, Davis*
Peng DING, *University of California, Berkeley*

Many scientific questions in biomedical research, environmental sciences, and psychology involve understanding the impact of multiple factors on an outcome of interest. Randomized factorial experiments are a popular tool for evaluating the causal effects of multiple treatments and their interactions simultaneously. However, drawing reliable causal inferences for multiple treatments in observational studies remains challenging. As the number of treatment combinations grows exponentially with the number of treatments, some treatment combinations can be rare or unobserved, posing additional difficulties in factorial effects estimation. To address these grand challenges, we propose a novel weighting approach tailored for observational studies with multiple treatments. Our approach uses weighted observational data to approximate a randomized factorial experiment, enabling us to estimate the effects of multiple treatments and their interactions simultaneously using the same set of weights. Our investigations suggest that the weights must balance the observed covariates and treatments for each contrast to provide unbiased estimates of the factorial effects of interest. Moreover, we discuss how to extend the proposed weighting method when some treatment combination groups are empty. Finally, we study the asymptotic behavior of the new weighting estimators and propose a consistent variance estimator, allowing researchers to conduct inferences for the factorial effects. Our approach is practical and widely applicable to various observational studies, providing a valuable tool for investigators interested in estimating the causal effects of multiple treatments.

242. Bayesian Shrinkage Kernel Regression for joint selection of microbiome data

[03.M1.I27, (page 19)]

Liangliang ZHANG, *Case Western Reserve University*

Recent advances in next generation sequencing technologies have allowed an increasing number of microbiome studies to be performed. Motivated by the structural similarities of microbiome data, with respect to several statistical properties such as, high-dimensionality, compositional nature, excess zeros due to low sequencing depth or dropout, we set out to build a Bayesian Shrinkage Kernel Regression model of associating microbiome data to patient level outcomes. This model is able to detect an effect of the overall mixture of microbial compositions in a flexible non-parametric way. In contrast to linear models, the Kernel function considers complex interactions among individual variables, integrates the non-additive effect of each variable, and assembles all the pieces into similarities across samples. Here are two main innovations. To model the dependence structure between variables, we modeled the structural similarities by constructing a double-space Gaussian Kernel which is an analogue to Unifrac Beta diversity. To encourage a selection of microbial variables within the Kernel, we proposed to shrink the weight parameters within the Gaussian Kernel using horseshoe priors. The model provides us with improved uncertainty assessment both at the joint level and the individual level.

243 . Determining PET brain activity using a Bayesian spatial model In Alzheimer's disease

[03.M1.I27, (page 19)]

Lijun ZHANG, *Case Western Reserve University*

Among the many neuroimaging modalities, positron emission tomography (PET) provides direct regional assessment of, among others,

brain metabolism, cerebral blood flow, amyloid deposition—all quantities of interest in the characterization of Alzheimer's disease (AD). However, there are analytic challenges in identifying early indicators of AD from these high-dimensional imaging data sets, and it is unclear whether early indicators of AD are more likely to emerge in localized patterns of brain activity or in patterns of correlation between distinct brain regions. Early PET-based analyses of AD focused on alterations in metabolic activity at the voxel-level or in anatomically defined regions of interest. Other approaches, including seed-voxel and multivariate techniques, seek to characterize metabolic connectivity by identifying other regions in the brain with similar patterns of activity across subjects. Here, we present an approach that provides a unified statistical framework for addressing both metabolic activity and connectivity. Specifically, we apply a Bayesian spatial hierarchical framework to longitudinal metabolic PET scans from the Alzheimer's Disease Neuroimaging Initiative.

244. Accounting for the spatial structure of weather systems in detected changes in precipitation extremes

[01.M1.I4, (page 4)]

Likun ZHANG, *University of Missouri*
Mark RISSER, *Lawrence Berkeley National Laboratory*
Edward MOLTER, *University of California, Berkeley*
Michael WEHNER, *Lawrence Berkeley National Laboratory*

The detection of changes over time in the distribution of precipitation extremes is complicated by noise at the spatial scale of weather systems. Traditional approaches for quantifying observed changes in extreme precipitation return values are often based on single-station analyses, which fail to account for the spatial coherence of individual storms and hence yield unrealistic and potentially misleading estimates of the true underlying changes in extremes. In this paper, we demonstrate how the use of a flexible statistical method that robustly accounts for the so-called "storm dependence" in measurements of daily precipitation removes a challenging source of noise and results in improved estimates of changes in the distribution of precipitation extremes. Furthermore, our analysis provides important insights into the spatial structure of seasonal extreme precipitation across increasing event rarity. Applying the methodology to long-term in situ records of daily precipitation from the central United States, we find that properly accounting for storm dependence leads to increased detection of statistically significant changes in return values as compared with existing approaches. We also find that simultaneous precipitation extremes in this region tend to organize on scales of 100-200 km for high quantile levels, which is consistent with observed spatial patterns in the NEXRAD Stage IV radar-based data set.

245. Optimizing Treatment Allocation in Randomized Clinical Trials by Leveraging Baseline Covariates

[01.E1.I23, (page 13)]

Zhiwei ZHANG, *Gilead Sciences*
Wei ZHANG, *Chinese Academy of Sciences*
Aiyi LIU, *National Institutes of Health*

We consider the problem of optimizing treatment allocation for statistical efficiency in randomized clinical trials. Optimal allocation has been studied previously for simple treatment effect estimators such as the sample mean difference, which are not fully efficient in the presence of baseline covariates. More efficient estimators can be obtained by incorporating covariate information, and modern machine learning methods make it increasingly feasible to approach full efficiency. Accordingly, we derive the optimal allocation ratio by maximizing the design efficiency of a randomized trial, assuming that an efficient estimator will be used for analysis. We then expand the scope of optimization by considering covariate-dependent randomization (CDR), which has some flavor of an observational study but provides the same level of scientific rigor as a standard randomized trial. We describe treatment effect estimators that are consistent, asymptotically normal and (nearly) efficient under CDR, and derive the optimal propensity score by maximizing the design efficiency of a CDR trial (under the assumption that an efficient estimator will be used for analysis). Our optimality results translate into optimal designs that improve upon standard practice. Real world examples and simulation results demonstrate that the proposed designs can produce substantial efficiency improvements in realistic settings.

246. Probability-of-Decision Designs to Accelerate Dose-Finding Trials

[01.A1.I14, (page 9)]

Tianjian ZHOU, *Colorado State University*

Cohort-based enrollment can slow down phase I dose-finding trials since the outcomes of the previous cohort must be fully evaluated before the next cohort can be enrolled. This results in frequent suspension of patient enrollment. We propose a class of probability-of-decision (POD) designs to accelerate dose-finding trials, which enable dose assignments in real-time in the presence of pending toxicity outcomes. With uncertain outcomes, the dose assignment decisions are treated as random variables, and we calculate the posterior distribution of the decisions. The posterior distribution reflects the variability in the pending outcomes and allows a direct and intuitive evaluation of the confidence of all possible decisions. Optimal decisions are calculated based on the 0-1 loss, and extra safety rules are constructed to enforce sufficient protection from exposing patients to risky doses. A new and useful feature of POD designs is that they allow investigators and regulators to balance the trade-off between enrollment speed and making risky decisions by tuning a pair of intuitive design parameters. The performances of POD designs are evaluated through numerical studies.

Directory

ABOWD, John
Cornell University and U.S. Census Bureau

Speaker: Plenary Lecture 2, p. 15, §1, p. 39

ACHARYYA, Suddhasatta
Gilead Sciences
SUDDHO_A@MSN.COM
Speaker: 03.A2.I42, p. 26, §2, p. 39

AGRAWAL, Shubhada
Georgia Institute of Technology
shubhadaiitd@gmail.com
Speaker: 03.A2.I41, p. 26, §3, p. 39

ALBERT, Jeffrey
Case Western Reserve University
jma13@case.edu
Speaker: 01.M1.I5, p. 4, §4, p. 39

ALBERT, Paul
National Cancer Institute

Speaker: Plenary Lecture 4, p. 17, §5, p. 40

ALT, Ethan
University of North Carolina at Chapel Hill
ethanalt@live.unc.edu
Speaker: 03.M1.I26, p. 19, §6, p. 40

AMONA, Elizabeth B
Virginia Commonwealth University
amonaeb@vcu.edu
Speaker: Student Paper Competition 1, p. 5, §7, p. 40

ANCESCHI, Niccolo
Duke University
niccolo.anceschi@duke.edu
Speaker: 01.M1.I6, p. 5, §8, p. 41

ANIM BEDIAKO, Theophilus
South Dakota State University
theophilus.animbediako@jacks.sdstate.edu

Speaker: Poster Session, p. 15, §9, p. 41

ARROYO, Jesús
Texas A&M University
jarroyo@tamu.edu
Speaker: 03.A1.I38, p. 24, §10, p. 41

ARYA, Sakshi
Pennsylvania State University
ska5950@psu.edu
Chair and organizer: 04.M1.I54, p. 32,
Speaker: 04.M1.I54, p. 32, §11, p. 42

ATHREYA, Avanti
Johns Hopkins University
dathrey1@jhu.edu
Chair and organizer: 03.M1.I30, p. 20,
Chair and organizer: 03.A1.I38, p. 24,
Speaker: 01.M2.I9, p. 6, §12, p. 42

AUSTERN, Morgane
Harvard
morgane.austern@gmail.com
Speaker: 03.A2.I45, p. 27, §13, p. 42

BAGCHI, Pramita
Department of Statistics, George Mason University
pbagchi@gmu.edu
Speaker: 03.M2.I33, p. 22, §14, p. 42

BAILEY, Maggie, D
Colorado School of Mines
mdbailey@mines.edu
Speaker: Poster Session, p. 15, §15, p. 43

BALASUBRAMANIAN, Krishna
UC Davis
kbala@ucdavis.edu
Chair: 03.E1.I46, p. 28,
Speaker: 03.E1.I46, p. 28, §16, p. 43

BANDYOPADHYAY, Soutir
Department of Applied Mathematics and Statistics, Colorado School of Mines
sbandyopadhyay@mines.edu
Chair: 01.A2.I17, p. 11,
Chair: Plenary Lecture 1, p. 15,
Speaker: 01.A2.I17, p. 11, §17, p. 43

BANERJEE, Anjishnu
Medical College of Wisconsin
abanerjee@mcw.edu
Chair: 01.A1.I16, p. 10,
Speaker: 01.M1.I1, p. 3, §18, p. 44

BANERJEE, Chitrak
Wells Fargo Bank, NA
chitrak.banerjee@wellsfargo.com
Speaker: 04.M2.I59, p. 34, §19, p. 44

BANERJEE, Hiya
Eli Lilly

Chair: Special Invited Session 1, p. 8,
Panelist: Panel Discussion 1, p. 10

BANERJEE, Paromita
JOHN CARROLL UNIVERSITY
pbanerjee@jcu.edu
Chair: 04.M1.I55, p. 32,
Chair: 04.M2.I60, p. 34,
Speaker: 04.M1.I55, p. 32, §20, p. 44

BANERJEE, Sayan
University of North Carolina, Chapel Hill
sayan@email.unc.edu
Speaker: 03.A1.I35, p. 24, §21, p. 45

BASAK, Piyali
Merck & Co.
pb15d@my.fsu.edu
Chair: 01.A1.I15, p. 9,
Chair and organizer: 01.A2.I20, p. 12,
Speaker: 01.M1.I6, p. 5, §22, p. 45

BASU, Sanjib
UIC
sbasu@uic.edu
Chair: 04.M1.I51, p. 31,
Speaker: 04.M1.I51, p. 31, §23, p. 45

BASU, Saonli
University of Minnesota
saonli@umn.edu
Chair: Memorial Session 3, p. 28

BAUGH, Samuel
Lawrence Berkeley National Laboratory
samuelbaugh@lbl.gov
Speaker: 03.A1.I37, p. 24, §24, p. 45

BEN-DAVID, Emanuel
U.S. Census Bureau

Speaker: Short Course 3, p. 31

BERA, Sabyasachi
University of Minnesota
berax008@umn.edu
Speaker: Poster Session, p. 15, §25, p. 45

BERESOVSKY, Vladislav
Bureau of Labor Statistics
beresovsky.vladislav@bls.gov
Speaker: 01.A2.I18, p. 11, §26, p. 46

BERG, Emily
Iowa State University
emilyb@iastate.edu
Speaker: 04.M2.I58, p. 33, §27, p. 46

BHADRA, Anindya
Purdue University
bhadra@purdue.edu
Speaker: 01.E1.I21, p. 13, §28, p. 46

BHADRA, Subhankar
North Carolina State University
sbhadra@ncsu.edu
Speaker: Poster Session, p. 15, §29, p. 47

BHATTACHARJEE, Monika
Indian Institute of Technology, Bombay

Organizer: 04.M2.I59, p. 33

BHATTACHARJEE, Samayita
University of California, Davis
saabhattacharjee@ucdavis.edu
Speaker: 03.M2.C1, p. 23, §30, p. 47

BHATTACHARYA, Anirban
Texas A&M University
anirbanb@stat.tamu.edu
Speaker: 03.A1.I40, p. 25, §31, p. 47

BHATTACHARYA, Anwesha
Wells Fargo

Speaker: Short Course 2, p. 23

BHATTACHARYYA, Arinjita
Merck & Co., Inc.
arinjita.bhattacharyya@merck.com

Organizer: 01.A1.I15, p. 9

BHATTACHARYYA, Sharmodeep
Oregon State University
sharmodeep.bhattacharyya@oregonstate.edu

Speaker: 03.E1.I50, p. 29, §32, p. 47

BHUYAN, Rashmi Ranjan
University of Southern California
bhuyanr@usc.edu
Speaker: Poster Session, p. 15, §33, p. 48

BORCHERT, Dylan, D
South Dakota State University
dylan.borchert@jacks.sdstate.edu

Speaker: Poster Session, p. 15, §34, p. 48

BOSE, Arup
Indian Statistical Institute
bosearu@gmail.com
Speaker: Memorial Session 3, p. 28, §35, p. 49

BRADLEY, Jonathan
Florida State University
jrbradley@fsu.edu
Speaker: 03.E1.I49, p. 29, §36, p. 49

BREIDT, Jay
NORC at the University of Chicago
jbreidt@gmail.com
Speaker: 01.M1.I3, p. 3, §37, p. 49

BRETZ, Frank
Novartis
frank.bretz@novartis.com
Speaker: Special Invited Session 4, p. 18, §38, p. 50

BRUCE, Scott
Texas A&M University
sabruce@tamu.edu
Speaker: 03.M2.I33, p. 22, §39, p. 50

BU, Fan
UCLA
fanbu@ucla.edu
Speaker: 01.M1.I1, p. 3, §40, p. 50

BUDHIRAJA, Amarjit
University of North Carolina Chapel Hill

Speaker: Bahadur Memorial Lecture, p. 17, §41, p. 51

BUKKE, Priyanjali
George Mason University
pbukke@gmu.edu
Speaker: Short Course 3, p. 31,
Speaker: Poster Session, p. 15, §42, p. 51

CALDER, Catherine
University of Texas at Austin
calder@austin.utexas.edu
Speaker: Special Invited Session 1, p. 8, §43, p. 51

CAO, Jian
Texas A&M University
jian.cao@tamu.edu
Speaker: 01.M2.I10, p. 7, §44, p. 51

CAPE, Joshua
University of Wisconsin
jrcape@wisc.edu
Speaker: 03.A1.I35, p. 23, §45, p. 52

CELENTANO, Michael
UC Berkeley
mcelentano@berkeley.edu
Speaker: 03.E1.I46, p. 28, §46, p. 52

CHAKRABARTY, Sayan
University of Illinois at Urbana Champaign
sayanc3@illinois.edu
Speaker: Poster Session, p. 16, §47, p. 52

CHAKRABORTY, Abhisek
Texas A&M University
abhisek_chakraborty@tamu.edu
Speaker: Student Paper Competition 1, p. 5, §48, p. 53

CHAKRABORTY, Nilanjan
Washington University in Saint Louis
chakra46@msu.edu
Chair: 03.M2.I34, p. 22,
Speaker: 01.A1.I12, p. 8, §49, p. 53

CHAKRABORTY, Sounak
University of Missouri
chakrabortys@missouri.edu
Speaker: 04.M1.I55, p. 33, §50, p. 53

CHAKRABORTY, Swarnita
Washington State University
s.chakraborty4@wsu.edu
Organizer: 01.E1.I24, p. 14

CHAN, Ivan
Bristol Myers Squibb

Panelist: Panel Discussion 2, p. 21

CHANDA, Aleena
University of Nebraska-Lincoln
achanda2@huskers.unl.edu
Speaker: Poster Session, p. 16, §51, p. 53

CHAPPELL, Rick
University of Wisconsin
chappell@stat.wisc.edu
Organizer: 01.E1.I22, p. 13,
Speaker: 01.M2.I8, p. 6, §52, p. 53,
Speaker: 01.E1.I22, p. 13, §53, p. 54

CHATLA, Suneel Babu
University of Texas at El Paso, Texas
chatla.suneel@gmail.com
Speaker: 03.M2.C1, p. 23, §54, p. 54

CHATTERJEE, Ansu
University of Minnesota
chatt019@umn.edu
Chair: Special Session 1, p. 6,
Organizer: 01.A2.I18, p. 11,
Chair: Plenary Lecture 3, p. 17,
Organizer: 03.M1.I28, p. 19,
Organizer: 04.M1.I51, p. 31,
Organizer: 04.M1.I52, p. 31,
Chair and organizer: 04.M2.I58, p. 33,
Speaker: 04.M1.I52, p. 31, §55, p. 54

CHATTERJEE, Sabyasachi
University of Illinois at Urbana Champaign
sc1706@illinois.edu
Organizer: 03.A2.I45, p. 27,
Speaker: 03.A2.I45, p. 27, §56, p. 54

CHATTERJEE, Shirshendu
City University of New York
shirchat1@gmail.com
Chair and organizer: 03.E1.I50, p. 29,
Speaker: 03.E1.I50, p. 29, §57, p. 54

CHATTOPADHYAY, Abhijnan
Postdoctoral Fellow, National Institute of Environmental Health Science, National Institute of Health
abhijnan.chattopadhyay@nih.gov
Speaker: 03.M2.C1, p. 23, §58, p. 54

CHAUDHURI, Sanjay
University of Nebraska-Lincoln
schaudhuri2@unl.edu
Chair: Special Invited Session 3, p. 12,
Chair: Special Invited Session 5, p. 21,
Speaker: 03.A2.I44, p. 27, §59, p. 55

CHEN, Sixia
University of Oklahoma Health Sciences Center
Sixia-Chen@ouhsc.edu
Speaker: 01.M1.I3, p. 4, §60, p. 55

CHEN, Yuting
University of Maryland, College Park
ychen215@umd.edu
Speaker: Poster Session, p. 16, §61, p. 55

COLUNGA, Elizabeth Juarez
GRECC

Panelist: Panel Discussion 1, p. 10

DAI, Fan
Michigan Technological University
fand@mtu.edu
Speaker: 03.A1.I39, p. 25, §62, p. 55

DANIELS, William, S
Colorado School of Mines
wdaniels@mines.edu
Speaker: Poster Session, p. 16, §63, p. 56

DAO, Mai
Wichita State University
mai.dao@wichita.edu
Speaker: 04.M2.I57, p. 33, §64, p. 56

DAS, Anisha
Florida State University
ad20fx@fsu.edu
Speaker: Poster Session, p. 16, §65, p. 56

DAS, Priyam
Virginia Commonwealth University
Priyam.Das@vcuhealth.org
Chair: 03.M1.I27, p. 19,
Speaker: 04.M1.I55, p. 32, §66, p. 56

DAS, Snigdha
Department of Statistics, Texas A&M University
snigdha@stat.tamu.edu
Speaker: Poster Session, p. 16, §67, p. 57

DAS, Soumojit
University of Maryland, College Park
soumojit@umd.edu
Speaker: Poster Session, p. 16, §68, p. 57

DASGUPTA, Nairanjana
Washington State University
dasgupta@wsu.edu
Panelist: Panel Discussion 1, p. 10,
Organizer: Panel Discussion 1, p. 10,
Chair: 01.E1.I24, p. 14,
Speaker: 01.E1.I24, p. 14, §69, p. 58

DATTA, Gauri
Univ of Georgia and US Census Bureau
gaurisdatta@gmail.com
Chair: 03.A2.I44, p. 27,
Speaker: 03.A2.I44, p. 27, §70, p. 58

DE, Simion
Biostatistics PhD Student, University of Minnesota
de000008@umn.edu
Speaker: 03.M2.C1, p. 22, §71, p. 58

DEMISSIE, Alemayehu
VP & Head of the Statistical Research and Data Science Center at Pfizer Inc.
demissie.alemayehu@pfizer.com
Chair: Special Session 2, p. 18

DESHPANDE, Sameer
University of Wisconsin–Madison
sameer.deshpande@wisc.edu
Speaker: 01.M1.I5, p. 5, §72, p. 58

DEWASKAR, Miheer
Duke University
miheer.dewaskar@duke.edu
Chair: 01.M1.I6, p. 5,
Speaker: 01.A2.I20, p. 12, §73, p. 59

DEY, Dipak
University of Connecticut
dipak.dey@uconn.edu
Speaker: 04.M1.I51, p. 31, §74, p. 59

DEY, Jyotirmoy
Regeneron Pharmaceuticals, Inc.
deyspeakable@gmail.com
Speaker: 03.A2.I42, p. 26, §75, p. 59

DEY, Tanujit
Harvard Medical School
tanujit.dey@gmail.com
Organizer: 01.M1.I5, p. 4,
Organizer: 01.A1.I16, p. 10,
Organizer: 03.M1.I27, p. 19,
Organizer: 04.M1.I55, p. 32

DHARAN, Bharani
Novartis Pharmaceuticals
jbdharan73@gmail.com
Chair and organizer: 01.A1.I14, p. 9,
Chair: Special Invited Session 2, p. 10,
Chair and organizer: 01.E1.I23, p. 13,
Chair: Special Invited Session 4, p. 18

DIAO, Guoqing
Milken Institute School of Public Health, George Washington University
yihuang@umbc.edu
Organizer: 03.M2.I33, p. 22

DUAN, Yaqi
MIT
yaqid@mit.edu
Speaker: 04.M2.I61, p. 34, §76, p. 60

DUBEY, Abhishek
Bristol Myers Squibb
abhishek.dubey@bms.com
Speaker: 01.E1.I23, p. 13, §77, p. 60

DUDEJA, Rishabh
Harvard University
rd2714@columbia.edu
Speaker: 03.E1.I46, p. 28, §78, p. 61

DUTTA, Diptavo
NIH/NCI
diptavo.dutta@nih.gov
Speaker: 01.A1.I13, p. 9, §79, p. 61

DUTTA, Somak
Iowa State University
somakd@iastate.edu
Organizer: 03.A1.I39, p. 25,
Speaker: 03.A1.I39, p. 25, §80, p. 62

ECKLUND, Dixie
University of Iowa
dixie-ecklund@uiowa.edu
Speaker: 01.E1.I22, p. 13, §81, p. 62

ERCIULESCU, Andreea
Westat
andreeaerciulescu@westat.com
Speaker: 01.A2.I18, p. 11, §82, p. 62

FOSDICK, Bailey
Colorado School of Public Health
bailey.fosdick@cuanschutz.edu
Speaker: 03.M1.I30, p. 20, §83, p. 62

FRANCO, Carolina
NORC at the University of Chicago
franco-carolina@norc.org
Speaker: 01.M1.I5, p. 4, §84, p. 63

GAILLIOT, Samuel F.
Texas A&M University
samuel.gailliot@stat.tamu.edu
Chair and organizer: 03.A1.I36, p. 24,
Speaker: Poster Session, p. 16, §85, p. 63

GANGULY, Indrila
North Carolina State University
igangul2@ncsu.edu
Speaker: Poster Session, p. 16, §86, p. 63

GAUSE, Christine
Merck

Panelist: Panel Discussion 2, p. 21

GHOSAL, Nairita
Merck & Co., Inc., Rahway, NJ, USA
nairita.ghosal@merck.com
Speaker: 01.A2.I20, p. 12, §87, p. 64

GHOSAL, Promit
Massachusetts Institute of Technology
promit@mit.edu
Speaker: 04.M2.I61, p. 35, §88, p. 64

GHOSAL, SATTWIK
IOWA STATE UNIVERSITY
ghosal@iastate.edu
Speaker: Poster Session, p. 16, §89, p. 64

GHOSH, Dhrubajyoti
Duke University
dg302@duke.edu
Speaker: 01.A1.I12, p. 8, §90, p. 64

GHOSH, Joyee
The University of Iowa
joyee-ghosh@uiowa.edu
Chair and organizer: 01.E1.I21, p. 13,
Speaker: 01.E1.I21, p. 13, §91, p. 65

GHOSH, Malay
University of Florida
ghoshm@stat.ufl.edu
Speaker: Memorial Session 1, p. 12, §92, p. 65

GHOSH, Souparno
University of Nebraska-Lincoln
soup.sigma@gmail.com
Chair and organizer: 04.M2.I57, p. 33,
Speaker: 04.M2.I57, p. 33, §93, p. 65

GHOSH, Tusharkanti
Colorado School of Public Health
TUSHARKANTI.GHOSH@CUANSCHUTZ.EDU

Chair: 03.M1.I31, p. 20,
Speaker: 03.M1.I31, p. 20, §94, p. 65

GHOSHAL, Subhashish
North Carolina State University
sghosal@ncsu.edu
Chair: Bahadur Memorial Lecture, p. 17

GIEFER, Clarissa L.
South Dakota State University
clarissa.giefer@jacks.sdstate.edu

Speaker: Poster Session, p. 16, §95, p. 65

GOH, Gyuhyeong
Kansas State University
ggoh@ksu.edu
Chair and organizer: Memorial Session 1, p. 12,
Organizer: 03.M2.I32, p. 21,
Speaker: 03.M2.I32, p. 22, §96, p. 66

GONG, Wenlong
University of Houston - Downtown
gongw@uhd.edu
Speaker: 03.A1.I37, p. 24, §97, p. 66

GREGORY, Karl
University of South Carolina
gregorkb@stat.sc.edu
Speaker: 01.A1.I12, p. 8, §98, p. 66

GUHA NIYOGI, Pratim
Johns Hopkins Bloomberg School of Public Health
pnyogi1@jhmi.edu
Speaker: 04.M2.I59, p. 34, §102, p. 67

GUHA, Aritra
AT&T Data Science and AI Research
aguha0109@gmail.com
Speaker: 01.A1.I13, p. 9, §99, p. 67

GUHA, Nilabja
University of Massachusetts Lowell
nilabja_guha@uml.edu
Chair and organizer: 03.A1.I40, p. 25,
Speaker: 03.A1.I40, p. 25, §100, p. 67

GUHA, Sharmistha
Texas A&M University
sharmistha@tamu.edu
Organizer: 01.M2.I11, p. 7,
Panelist: Panel Discussion 1, p. 10,
Speaker: 01.M2.I11, p. 7, §101, p. 67

GUHANIYOGI, Rajarshi
Texas A&M University
rajguhaniyogi@tamu.edu
Chair: 01.M2.I11, p. 7,
Chair and organizer: 01.A2.I19, p. 11,
Chair and organizer: 04.M1.I56, p. 33,
Speaker: 03.A1.I36, p. 24, §103, p. 68

GUINDANI, Michele
UCLA, Biostatistics
mguindani@g.ucla.edu
Speaker: 01.A2.I19, p. 12, §104, p. 68

GUINNESS, Joseph
Cornell University
guinness@cornell.edu
Speaker: 01.M2.I10, p. 7, §105, p. 68

GUNTUBOYINA, Adityanand
University of California Berkeley
aditya@stat.berkeley.edu
Chair and organizer: 01.M1.I2, p. 3,
Chair and organizer: 01.E1.I25, p. 14,
Chair: 03.A2.I45, p. 27,
Organizer: 03.E1.I48, p. 29,
Speaker: 03.E1.I48, p. 29, §106, p. 68

GUTIERREZ, Rene
Texas A&M University
renegutierrez@tamu.edu
Speaker: 03.A1.I36, p. 24, §107, p. 68

GWON, Yeongjin
University of Nebraska Medical Center
yeongjin.gwon@unmc.edu
Speaker: 03.M2.I32, p. 21, §108, p. 68

HARRIS, Trevor
Texas A&M University
tharris@stat.tamu.edu
Chair: 01.M1.I4, p. 4,
Speaker: 01.M2.I10, p. 7, §109, p. 69

HASSE, Jason
South Dakota State University
jason.hasse@jacks.sdstate.edu
Speaker: Poster Session, p. 16, §110, p. 69

HELU, Amal
The University of Jordan
al_helu@yahoo.com
Speaker: 01.M2.I7, p. 6, §111, p. 69

HESTERBERG, Tim
Instacart
timhesterberg@gmail.com
Panelist: Panel Discussion 2, p. 21,
Speaker: 03.M1.I28, p. 19, §112, p. 69,
Speaker: Special Session 1, p. 6, §113, p. 69

HEYMAN, Megan
Rose-Hulman Institute of Technology
heyman@rose-hulman.edu
Chair: 03.M1.I28, p. 19,
Speaker: 03.M1.I28, p. 19, §114, p. 69

HOGUE, Terry
Dean, Earth and Society Programs, Colorado School of Mines

Speaker: Conference Inaugaration, p. 3

HOOTEN, Mevin
The University of Texas at Austin
mevin.hooten@austin.utexas.edu
Speaker: 04.M1.I56, p. 33, §115, p. 70

HORE, Gaurab
University of Maryland, Baltimore County
gaurabh1@umbc.edu
Speaker: Student Paper Competition 1, p. 5, §116, p. 70

HUANG, Yi
University of Maryland, Baltimore County
yihuang@umbc.edu
Chair and organizer: 03.M1.I29, p. 20,
Chair: 03.M2.I33, p. 22

HUSSEIN, Abdulkadir
University of Windsor
ahussein@uwindsor.ca
Speaker: Special Session 2, p. 18, §117, p. 70

LAHA, Nilanjana
Texas A&M
nilanjanaaa.laha@gmail.com
Chair and organizer: 04.M2.I61, p. 34

LAHIRI, Partha
University of Maryland College Park
plahiri@umd.edu
Chair: 04.M1.I52, p. 31

LAHIRI, Soumendranath
Washington university of St. Louis

Organizer: 01.A1.I12, p. 8,
Organizer: 03.M2.I34, p. 22,
Speaker: Memorial Session 2, p. 26, §142, p. 77

LARRY, Leon
Merck
larry.leon2@merck.com
Speaker: 03.M1.I26, p. 19, §143, p. 77

LE, Can
University of California, Davis
canle@ucdavis.edu
Speaker: 03.E1.I50, p. 29, §144, p. 78

LEVIN, Keith
University of Wisconsin-Madison
kdlevin@wisc.edu
Speaker: 03.M1.I30, p. 20, §145, p. 78

LI, YAN
UNIVERSITY OF MARYLAND AT COLLEGE PARK
YLI6@UMD.EDU
Speaker: 04.M2.I58, p. 33, §146, p. 78

LINERO, Antonio
The University of Texas at Austin
Antonio.Linero@austin.utexas.edu

Speaker: 01.A2.I20, p. 12, §147, p. 79

LIPKOVICH, Ilya
Eli Lilly and Company
ilya.lipkovich@lilly.com
Speaker: 01.A1.I14, p. 9, §148, p. 79

LOH, Po-Ling
University of Cambridge
polingloh@gmail.com
Organizer: 03.E1.I46, p. 28,
Organizer: 04.M1.I53, p. 32

LOH, Po-Shen
Carnegie Mellon University

Panelist: Special Session 1, p. 6

LUNDE, Robert
Washington University in St. Louis
lunde@wustl.edu
Speaker: 01.M2.I9, p. 7, §149, p. 79

MA, Wanying
Novartis Pharmaceuticals Corporation
wanying.ma@novartis.com
Speaker: 01.A1.I14, p. 9, §150, p. 79

MAITRA, Ranjan
Iowa State University
maitra@iastate.edu
Speaker: 03.A1.I39, p. 25, §151, p. 80

MAITY, Arnab
Pfizer
arnab.maity@pfizer.com
Chair and organizer: 01.M2.I7, p. 6,
Speaker: 01.A1.I15, p. 10, §152, p. 80

MAJUMDER, Reetam
NC State University
reetam.m303@gmail.com
Organizer: 01.M1.I4, p. 4,
Chair and organizer: 01.M2.I10, p. 7,
Speaker: 01.M1.I4, p. 4, §153, p. 80

MAK, Simon
Duke University
sm769@duke.edu
Speaker: 03.A2.I43, p. 26, §154, p. 81

MALEKI, Arian
Columbia University
mm4338@columbia.edu
Speaker: 03.A2.I45, p. 27, §155, p. 81

MALLICK, Himel
Cornell University
himel.stat.iitk@gmail.com
Organizer: 01.M1.I6, p. 5

MANATUNGA, Amita
Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University

Speaker: Special Invited Session 2, p. 10, §156, p. 81

MANDAL, Soutrik
NYU Grossman School of Medicine
Soutrik.Mandal@nyulangone.org
Speaker: 03.E1.I47, p. 28, §157, p. 82

MAZUMDER, Rahul
MIT Sloan School of Management
rahulmaz@mit.edu
Speaker: Special Invited Session 5, p. 21, §158, p. 82

MCELROY, Tucker
US Census Bureau
Tucker.S.McElroy@census.gov
Organizer: 01.A2.I17, p. 11

MELIKECHI, Omar
Duke University
omar.melikechi@duke.edu
Speaker: 01.M1.I6, p. 5, §159, p. 82

MEYER, Mary
Colorado State University
marycmeyer@gmail.com
Speaker: Special Invited Session 1, p. 8, §160, p. 82

MICHAEL, Semhar
South Dakota State University
semhar.michael@sdstate.edu
Speaker: Special Session 2, p. 18, §161, p. 83

MICHAILIDIS, George
UCLA
gmichail@ucla.edu
Speaker: Special Invited Session 5, p. 21, §162, p. 83

MISRA, Neeraj
Indian Institute of Technology, Kanpur, India
neeraj@iitk.ac.in
Chair: 04.M2.I59, p. 33,
Speaker: 04.M2.I59, p. 34, §163, p. 83

MOLLOY, R. Cole
JHU/APL
Robert.Molloy@jhuapl.edu
Speaker: 03.A2.I43, p. 27, §164, p. 83

RASHID, Naim
Department of Biostatistics, Gillings School of Global Public Health, UNC-CH
naim@unc.edu
Speaker: 01.A1.I16, p. 10, §190, p. 91

ROY CHOWDHURY, Niladri
Bristol-Myers-Squibb
niladri.roychowdhury@bms.com
Speaker: 01.A1.I15, p. 9, §194, p. 92

ROY, Anindya
University of Maryland Baltimore County
anindya@umbc.edu
Chair: Plenary Lecture 2, p. 15,
Chair: Plenary Lecture 4, p. 17,
Organizer: 03.M1.I31, p. 20,
Chair: Memorial Session 2, p. 26

ROY, Arkaprava
University of Florida
ark007@ufl.edu
Chair and organizer: 01.A1.I13, p. 9,
Speaker: 01.A2.I19, p. 12, §191, p. 91

ROY, Asmita
Texas A&M University
asmita112358@tamu.edu
Speaker: 03.M2.C1, p. 22, §192, p. 92

ROYCHOUDHURY, Satrajit
Pfizer Inc.
satrajit.roychoudhury@pfizer.com
Speaker: Special Invited Session 4, p. 18, §193, p. 92

RUSSELL, Brook
Clemson University School of Mathematical and Statistical Sciences
brookr@clemson.edu
Speaker: 01.M1.I4, p. 4, §195, p. 93

SADHU, Ritwik
Department of Statistics and Data Science, Cornell.University
rs2526@cornell.edu
Speaker: Student Paper Competition 2, p. 7, §196, p. 93

SAEGUSA, Takumi
University of Maryland
tsaegusa@umd.edu
Speaker: 03.M2.I32, p. 21, §197, p. 93

SAHA, Abhijoy
Eli Lilly and Company
saha_abhijoy@lilly.com
Chair: 03.A2.I43, p. 26,
Speaker: 01.E1.I23, p. 14, §198, p. 94

SAHA, Arkajyoti
University of Washington, Department of Statistics
arkajyotisaha93@gmail.com
Speaker: 03.A1.I37, p. 24, §199, p. 94

SAHA, Satabdi
The University of Texas MD Anderson Cancer Center
ssaha1@mdanderson.org
Speaker: 03.E1.I47, p. 28, §200, p. 94

SAHOO, Indranil
Virginia Commonwealth University
sahooi@vcu.edu
Chair and organizer: 03.A1.I37, p. 24

SANG, Huiyan
Texas A&M University
hs37@tamu.edu
Speaker: 04.M1.I51, p. 31, §201, p. 95

SARKAR, Abhra
The University of Texas at Austin
abhra.stat@gmail.com
Speaker: 04.M2.I60, p. 34, §202, p. 95

SARKAR, Purnamrita
University of Texas at Austin
purna.sarkar@austin.utexas.edu
Speaker: Special Invited Session 6, p. 23, §203, p. 95

SAVITSKY, Terrance
BLS
Savitsky.Terrance@bls.gov
Speaker: 01.A2.I18, p. 11, §204, p. 96

SEN, Aditi
University of Maryland
asen123@umd.edu
Speaker: Poster Session, p. 17, §205, p. 96

SEN, Bodhisattva
Columbia University
bodhisattva.sen@gmail.com
Chair: 03.E1.I48, p. 29,
Speaker: 01.E1.I25, p. 14, §206, p. 96

SEN, Subhabrata
Harvard University
subhabratasen@fas.harvard.edu
Speaker: 03.A1.I38, p. 25, §207, p. 96

SENGUPTA, Srijan
North Carolina State University
ssengup2@ncsu.edu
Chair and organizer: 01.M1.I1, p. 3,
Chair and organizer: 01.M2.I9, p. 6,
Speaker: Short Course 1, p. 8,
Chair and organizer: 03.A1.I35, p. 23,
Chair and organizer: 03.E1.I49, p. 29,
Speaker: 03.M1.I30, p. 20, §208, p. 96

SETHURAMAN, Sunder
University of Arizona
sethuram@math.arizona.edu
Speaker: Memorial Session 3, p. 28, §209, p. 97

SHABY, Ben
Colorado State University
bshaby@colostate.edu
Speaker: 04.M1.I56, p. 33, §210, p. 97

SHAH, Harsh Hemant
University of Minnesota
shah0830@umn.edu
Speaker: Poster Session, p. 17, §211, p. 97

SHARMA, Gaurav
Takeda Pharmaceuticals
gaurav.sharma2@takeda.com
Speaker: 03.M1.I29, p. 20, §212, p. 97

SHARMA, Somya
University of Minnesota - Twin Cities
sharm636@umn.edu
Speaker: Student Paper Competition 1, p. 5, §213, p. 97

YE, Jiabu
Merck Research Labs
jiabu.ye@merck.com
Chair: 01.E1.I22, p. 13,
Speaker: 01.M2.I8, p. 6, §239, p. 104

YU, Cheng-Han
Marquette University
cheng-han.yu@marquette.edu
Speaker: 03.A1.I36, p. 24, §240, p. 104

YU, Ruoqi
University of California, Davis
ruoqi.yu.ry@gmail.com
Speaker: 01.M1.I2, p. 3, §241, p. 105

ZHANG, Liangliang
Case Western Reserve University
lxz716@case.edu
Speaker: 03.M1.I27, p. 19, §242, p. 105

ZHANG, Lijun
Case Western Reserve University
lxz759@case.edu
Speaker: 03.M1.I27, p. 19, §243, p. 105

ZHANG, Likun
University of Missouri
likun.zhang@missouri.edu
Speaker: 01.M1.I4, p. 4, §244, p. 106

ZHANG, Zhiwei
Gilead Sciences
Zhiwei.Zhang6@gilead.com
Speaker: 01.E1.I23, p. 13, §245, p. 106

ZHOU, Tianjian
Colorado State University
tianjian.zhou@colostate.edu
Speaker: 01.A1.I14, p. 9, §246, p. 106