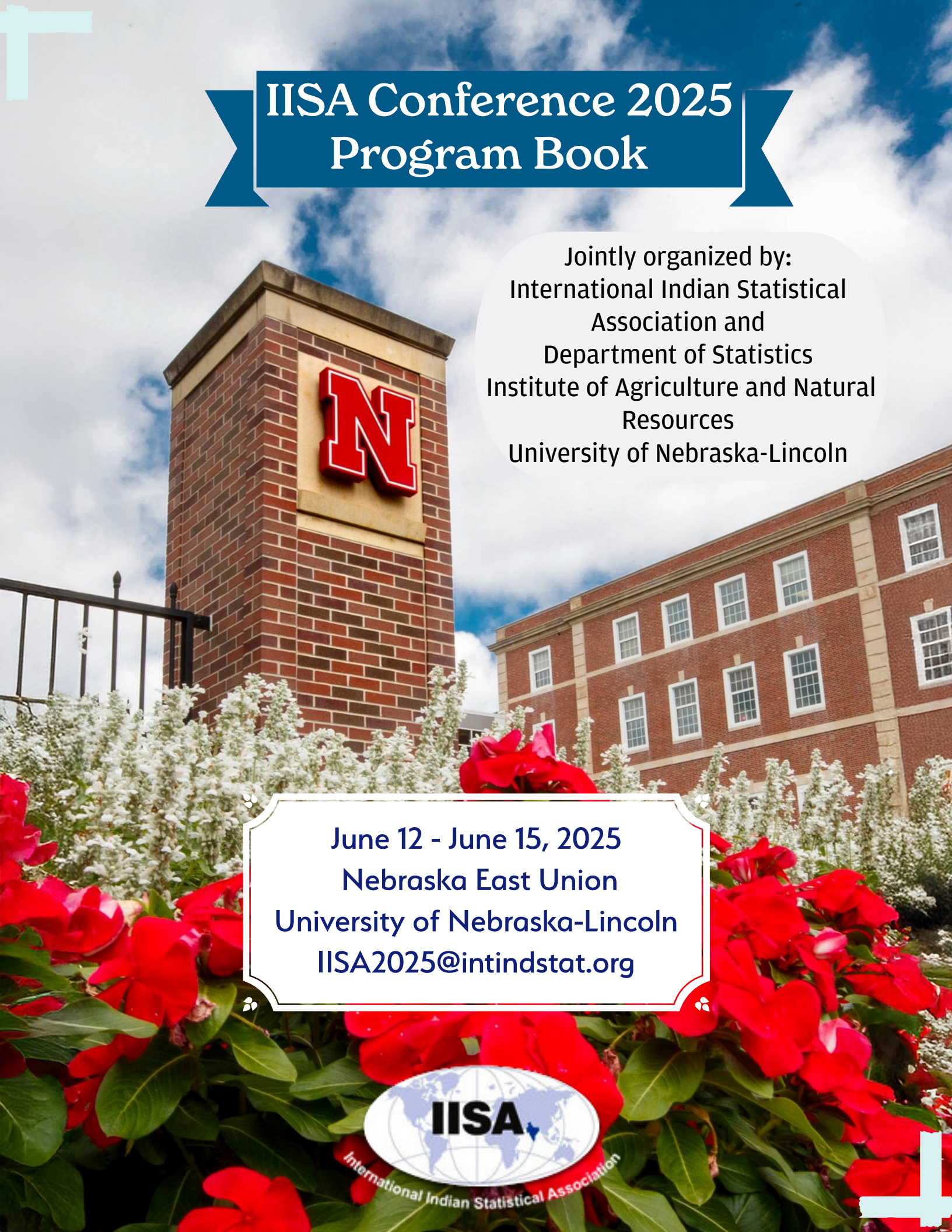
IISA Conference 2025 Program Book



Jointly organized by:
International Indian Statistical Association and
Department of Statistics
Institute of Agriculture and Natural Resources
University of Nebraska-Lincoln

June 12 - June 15, 2025
Nebraska East Union
University of Nebraska-Lincoln
IISA2025@intindstat.org

IISA
International Indian Statistical Association

International Indian Statistical Association

Annual Conference

Nebraska East Union

University of Nebraska-Lincoln

Lincoln, Nebraska, USA

12th - 15th June, 2025

Scientific Programme Committee
Chair: Bodhisattva SEN, Columbia University, USA
Co-chair: Po Ling LOH, Cambridge University, UK
Co-chair: Margaret GAMALO, Pfizer, USA

Members:
Marco AVELLA MEDINA, Columbia University, USA
Sayantan BANERJEE, IIM Indore, India
Piyali BASAK, Merck, USA
Abhishek BHATTACHARJEE, Pfizer, USA
Anirban BHATTACHARYA, Texas A&M University, USA
Bhaswar BHATTACHARYA, University of Pennsylvania, USA
Arup BOSE, ISI Calcutta, India
Hao CHEN, University of California Davis, USA
Tirthankar DASGUPTA, Rutgers University, USA
Abhirup DATTA, Johns Hopkins University, USA
Nabarun DEB, University of Chicago, USA
Souvik DHARA, Purdue University, USA
Peng DING, University of California Berkeley, USA
Karin S DORMAN, Iowa State University, USA
Paromita DUBEY, University of Southern California, USA
Sarmistha GUHA, Texas A&M University, USA
Adityanand GUNTUBOYINA, University of California Berkeley, USA
Dong-yun KIM, NIH, USA
Madan KUNDU, Daiichi Sankyo, USA
Jianchang LIN, Takeda, USA
Ashwin PANANJADY, Georgia Tech University, USA
Debdeep PATI, University of Wisconsin Madison, USA
Rohit PATRA, LinkedIn, USA
Himel MALLICK, Cornell University, USA
Neeraj MISHRA, IIT Kanpur, India
Abhyuday MANDAL, University of Georgia, USA
Rajesh NAIR, FDA, USA
Garvesh RASKUTTI, University of Wisconsin Madison, USA
Cindy RUSH, Columbia University, USA
Ksheera SAGAR, Corteva, USA
Purnamrita SARKAR, University of Texas at Austin, USA
Rajen SHAH, Cambridge University, USA
Subhabrata SEN, Harvard University, USA
Li WANG, Abbvie, USA
Miaoyan WANG, University of Wisconsin Madison, USA

Local Organising Committee
Chair: Bertrand CLARKE, University of Nebraska-Lincoln
Sanjay CHAUDHURI, University of Nebraska-Lincoln
Su CHEN, University of Nebraska Medical Center
Jennifer CLARKE, University of Nebraska-Lincoln
Souparno GHOSH, University of Nebraska-Lincoln
Heike HOFFMAN, University of Nebraska-Lincoln
Reka HOWARD, University of Nebraska-Lincoln
Indranil MUKHOPADHYAY, University of Nebraska-Lincoln

International Organizing Committee:
Saonli BASU
Hiya BANERJEE
Snigdhansu CHATTERJEE
Sanjay CHAUDHURI
Jyotiska DATTA
Subrata KUNDU
Abhyuday MANDAL
Arkaprava ROY
Ananda SEN

Student and Staff Volunteers:
Asif ENAN, Oluwafunmibi FASANYA, Maksuda Aktar TOMA, Xeng YANG, Kalani Hasanthika PAHALAPATHIRAGE DONA, Dinuwanthi LIYANAGE, Malith PREMARATHNA, Enakshy DUTTA, Gayara Demini F MUTHUNAMA GONNAGE, Riddhimoy GHOSH, Aftab A SORWAR, Aleena CHANDA, Beenu SAREENA, Arian ALAI, Muxin HUA.

Sponsorships and Endorsements:

From The President, IISA

On behalf of the International Indian Statistical Association (IISA), it is my great pleasure to welcome you to the IISA 2025 Annual Conference. This gathering marks the flagship event for our organization, providing a unique platform for statisticians, data scientists, and researchers from across the globe to share ideas, forge collaborations, and celebrate the progress of our profession.

The IISA is a vibrant, inclusive, and globally connected non-profit organization, with membership open to all. Our core objectives are:

- To promote education, research, and application of statistics, probability, and data science worldwide—with a special emphasis on the Indian subcontinent.
- To foster international collaboration in advancing statistical science.
- To support and empower early career statisticians and data scientists.
- To encourage synergy between academia, industry, and government sectors in the statistical and data science communities.

Following the tradition of past IISA conferences, IISA 2025 promises a rich scientific program that brings together distinguished speakers and young researchers to explore recent developments in statistical theory, methods, applications, and the expanding role of data science in evidence-based decision-making.

This year, the conference will host:

- Four Plenary Lectures,
- Ten Special Invited Talks,
- More than 74 Invited Sessions with over 230 presentations,
- Student paper and poster competitions featuring the work of rising stars in our field,
- Panel discussions and
- STATSBOWL competition

A big thank you to the University of Nebraska at Lincoln for so graciously hosting us this year! We're also incredibly grateful to the scientific program committee, the local organizing committee, student paper committee, session organizers, and all the amazing volunteers, this conference wouldn't be possible without your hard work and dedication.

We would like to extend our sincere thanks to our sponsors—the American Statistical Association (ASA), National Science Foundation (NSF), ASA Biopharmaceutical Section, Eli Lilly and Company, Merck, and Pfizer—for their generous support of this conference.

Thank you for joining us. We look forward to an engaging, insightful, and inspiring IISA 2025 conference experience with you.

Warm regards,

Hiya Banerjee
President, International Indian Statistical Association
Eli Lilly and Company

AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

Dear Colleagues, June, 2025

On behalf of the entire American Statistical Association community, I extend our warmest congratulations to the International Indian Statistical Association on hosting what promises to be an exceptional and impactful conference. Your dedication to advancing our shared mission and fostering professional excellence continues to inspire us all.

As the conference attendees gather, we are reminded that our individual associations are strongest when we work together toward common goals. The innovative program reflects that spirit: 4 plenary talks, 10 special invited talks, 2 short courses, and over 60 invited sessions. It's a strong lineup that shows what we can achieve through collaboration. Your commitment to "bringing together leading statisticians, data scientists, and researchers from around the world to discuss recent developments in theory, methods, and applications of statistics and data science" will undoubtedly generate new insights and forge lasting connections.

As president of the ASA, I believe deeply in building bridges between organizations, disciplines, and communities. When we strengthen these connections, we amplify our collective impact and create opportunities that extend far beyond what any single association could achieve alone. This conference is a perfect example of how thoughtful collaboration can elevate the entire community and advance the science we all care about.

I want to acknowledge Bodhisattva Sen, Bertrand Clarke, Hiya Banerjee, and all the members of the conference committees for creating a conference where our community can learn, grow, and thrive together. We are honored to support this conference, which will serve as a catalyst for continued cooperation and shared success.

With appreciation and best wishes for a successful conference.

Ji-Hyun Lee
2025 President

amstat.org asainfo@amstat.org

LINCOLN

June 11, 2025

Dear friends and members of the International Indian Statistical Association,

As Mayor of Lincoln, it is my distinct pleasure to welcome you on June 12 – June 15, 2025, for the International Indian Statistical Association (IISA) Conference at the University of Nebraska-Lincoln. We are delighted that you chose this city campus as your accommodation and feel confident you will have an inspired and productive experience that is also comfortable and enjoyable.

For those of you who are visiting, we know you'll find the City of Lincoln to be a warm host. Our citizens take great pride in our quality of life, hospitality, and friendliness. We invite you to take time to enjoy our many attractions, including the wonderful shopping and eateries throughout downtown and in the nearby Historic Haymarket and Railyard area. If you have time, there are some wonderful new exhibits in the recently expanded Lincoln Children's Zoo near the Sunken Gardens, or you can stroll towards the Antelope Valley Trail and Union Plaza Park for fresh air and relaxation.

Our sincere gratitude to each of you for your outstanding efforts in your crucial fields of work.

I hope you enjoy your stay in Lincoln, and that you will come back often!

Sincerely,

Leirion Gaylor Baird

Leirion Gaylor Baird
Mayor of Lincoln

CITY OF LINCOLN | OFFICE OF THE MAYOR | 555 SOUTH 10TH STREET, SUITE 301 | LINCOLN, NE 68508

Contents

Program Overview

Program

Thursday June 12

Conference Inaugaration

Venue: Plains A

Time : **08:30 - 09:00**

- **Hiya BANERJEE**, Eli Lilly and Company
- **Tiffany HENG-MOSS**, University of Nebraska-Lincoln
- **Bodhisattva SEN**, Columbia University
- **Bhaskar BHATTACHARYA**, University of Nebraska-Lincoln

Plenary Lecture 1 Debashis Ghosh

Venue: Plains A

Chair : Souparno GHOSH, University of Nebraska-Lincoln

09:15 **Reproducibility in Statistics** [Abstract 80]

Debashis GHOSH, *University of Colorado Anschutz Medical Campus*

Time : **10:15 - 10:30** **Coffee Break** *Venue:* Foyer

Panel Discussion 1 Women in Statistics: Breaking Barriers and Shaping the Future

Venue: Plains A

Time : **10:30 - 12:00**

Moderator: Hiya BANERJEE, Eli Lilly and Company
Sponsor: Caucas for Women in Statistics

- **Saonli BASU**, University of Minnesota
- **Fanni NATANEGARA**, Eli Lilly
- **Bhaskar BHATTACHARYA**, University of Nebraska-Lincoln
- **Susmita DATTA**, University of Florida

01.M2.I1 Interpretable Machine Learning

Venue: Plains B

Chair : Lili ZHENG, University of Illinois Urbana - Champaign
Organizer : Garvesh RASKUTTI, University of Wisconsin-Madison

10:30 **TBD** [Abstract 209]

Kris SANKARAN, *University of Wisconsin-Madison*

11:00 **TBD** [Abstract 181]

Gunwoong PARK, *Seoul National University*

11:30 **A model-agnostic ensemble framework with built-in LOCO feature importance inference** [Abstract 262]

Lili ZHENG, *University of Illinois Urbana - Champaign*
Luqin GAN,
Genevera ALLEN, *Columbia University*

01.M2.I2 Recent Developments in Statistical Inference

Venue: Plains C

Chair : Abhinav CHAKRABORTY, Columbia University

10:30 Multistage drop-the-losers designs for selecting the effective treatment(s) and estimating their worth [Abstract 158]

Neeraj MISRA, *Indian Institute of Technology Kanpur*
Yogesh KATARIYA, *Indian Institute of Technology Kanpur*

11:00 Minimax And Adaptive Transfer Learning for Nonparametric Classification under Distributed Differential Privacy Constraints [Abstract 30]

Abhinav CHAKRABORTY, *Columbia University*
Arnab AUDDY, *Ohio State University*
Tony CAI, *University of Pennsylvania*

01.M2.I3 Optimal Transport: Recent Advances in Machine Learning and Statistics

Venue: Prairie A

Chair and Organizer : Nilanjan CHAKRABORTY, Missouri University of Science and Technology

10:30 No-Regret Generative Modeling via Parabolic Monge-Ampère PDE [Abstract 62]

Nabarun DEB, *University of Chicago*
Tengyuan LIANG, *University of Chicago*

11:00 Approximation rates of entropic maps in semidiscrete optimal transport [Abstract 204]

Ritwik SADHU, *Amazon*
Ziv GOLDFELD, *Cornell University*
Kengo KATO, *Cornell University*

11:30 Geometric Exploration of Random Objects using Distance Profiles [Abstract 67]

Paromita DUBEY, *University of Southern California*
Yaqing CHEN, *Rutgers University*
Hans-Georg MÜLLER, *UC Davis*

01.M2.I4 Recent Development in Survival Analysis with Biomedical Applications

Venue: Prairie B

Chair : Yue ZHAN, University of Nebraska Medical Center
Organizer : Xiaotian CHEN, Abbvie

10:30 Assessing Contribution of Treatment Phases through Tipping Point Analyses via Counterfactual Elicitation Using Rank Preserving Structural Failure Time Models [Abstract 20]

Sudipta BHATTACHARYA, *Daiichi Sankyo, Inc.*
Jyotirmoy DEY, *Regeneron*

11:00 Exploring potential treatment effect heterogeneities in clinical trials with time to event endpoints [Abstract 153]

Alejandro MANTERO, *GSK*

11:30 Joint analysis for multivariate longitudinal and interval-censored event time data: Application in Huntington's disease [Abstract 258]

Yue ZHAN, *University of Nebraska Medical Center*
Cheng ZHENG, *University of Nebraska Medical Center*
Ying ZHANG, *University of Nebraska Medical Center*

Student Paper Competition 1 Probability and Theory of Statistics and Data Sciences

Venue: Prairie C
Chair : Snigdha PANIGRAHI, University of Michigan

10:30 Bayesian Semi-supervised Inference via a Debiased Modeling Approach [Abstract 216]

Gözde SERT, *Texas A&M University*
Gözde SERT, *Texas A&M University*
Abhishek CHAKRABORTTY, *Texas A&M University*
Anirban BHATTACHARYA , *Texas A&M University*

10:45 Signal-to-noise ratio aware minimax analysis of sparse linear regression [Abstract 84]

Shubhangi GHOSH, *Columbia University*
Yilin GUO,
Haolei WENG,
Arian MALEKI, *Columbia University*

11:00 Assumption-lean weak limit of weighted IPW estimator for two-stage adaptive experiments [Abstract 174]

Ziang NIU, *University of Pennsylvania*
Zhimei REN, *University of Pennsylvania*

11:15 Power properties of the two-sample test based on the nearest neighbors graph [Abstract 113]

Rahul Raphael KANEKAR, *Stanford University*

11:30 Bridging Root-nand Non-standard Asymptotics: Adaptive Inference in M-Estimation [Abstract 119]

Kenta TAKATSU, *Carnegie Mellon University*
Arun Kumar KUCHIBHOTLA, *Carnegie Mellon University*

Student Paper Competition 2 Application of Statistics and Data Sciences

Venue: Arbor
Chair : Satarupa BHATTACHARJEE, University of Florida

10:30 Aligning Multiple Inhomogeneous Random Graphs: Fundamental Limits of Exact Recovery [Abstract 4]

Taha AMEEN, *University of Illinois Urbana-Champaign*
Bruce HAJEK, *University of Illinois Urbana-Champaign*

10:45 Nonparametric Within-Between Models [Abstract 25]

Soumyabrata BOSE, *University of Texas at Austin*
Antonio R. LINERO, *University of Texas at Austin*
Jared S. MURRAY, *University of Texas at Austin*

11:00 Hypothesis selection via sample splitting for valid powerful testing in matched observational studies [Abstract 55]

Abhinandan DALAL, *University of Pennsylvania*
William BEKERMAN, *University of Pennsylvania*
Carlo DEL NINNO, *Formerly of the World Bank*
Dylan SMALL, *University of pennsylvania*

11:15 Risk-inclusive Contextual Bandits for Early Phase Clinical Trials [Abstract 115]

Rohit KANRAR, *Iowa State University*
Chunlin LI,

Zara GHODSI,
Margaret GAMALO,

11:30 Modeling Spatial Extremes using Non-Gaussian Spatial Autoregressive Models via Convolutional Neural Networks [Abstract 192]

Sweta RAI, *Colorado School of Mines*
Sweta RAI, *Colorado School of Mines*
Douglas NYCHKA, *Colorado School of Mines*
Soutir BANDYOPADHYAY, *Colorado School of Mines*

Time : **12:00 - 13:30 Lunch** *Venue:* **East Campus Dining Center**

Bahadur Memorial Lecture Sourav Chatterjee

Venue: **Plains A**

Chair : Subhashis GHOSHAL, North Carolina State University

13:30 Neural networks can learn any low complexity pattern [Abstract 45]

Sourav CHATTERJEE, *Stanford University*
Timothy SUDIJONO,

Time : **14:30 - 14:45 Short Break No Coffee**

Student Poster Competition

Venue: **Foyer**

Time : **14:45 - 18:00**

- **Inference on Network Structures of Hypergraphs under Edge Exchangeability** [Abstract 19]

  Ayoushman BHATTACHARYA, *Department of Statistics and Data Science, Washington University in St. Louis*
  Nilanjan CHAKRABORTY, *Department of Mathematics and Statistics, Missouri University of Science and Technology*
  Robert LUNDE, *Department of Statistics and Data Science, Washington University in St. Louis*

- **Performance Guaranteed Confidence Sets of Ranks** [Abstract 39]

  Onrina CHANDRA, *Rutgers University*
  Minge XIE,

- **Inference for projection parameters in Linear Regression** [Abstract 40]

  Woonyoung CHANG, *Carnegie Mellon University*
  Arun KUCHIBHOTLA, *Carnegie Mellon University*
  Alessandro RINALDO, *University of Texas at Austin*

- **PriME: Privacy-aware Membership profile Estimation in networks** [Abstract 42]

  Sayak CHATTERJEE, *University of Pennsylvania*
  Abhinav CHAKRABORTY, *University of Pennsylvania*
  Sagnik NANDY, *University of Chicago*

- **Bayesian Mechanistic Model of Spatio-Temporal Dynamics in Invasive Tree** [Abstract 49]

  Jiaqi CHEN, *University of Nebraska-Lincoln*
  Yawen GUAN,
  Huijing DU,

- **Scalable Efficient Inference in Complex Surveys through Targeted Resampling of Weights** [Abstract 57]

  Snigdha DAS, *Texas A&M University*
  Dipankar BANDYOPADHYAY, *Virginia Commonwealth University*
  Debdeep PATI, *University of Wisconsin - Madison*

- **Differentially Private Bayesian Tests** [Abstract 58]

  Saptati DATTA, *Texas A&M University*
  Abhisek CHAKRABORTY,

- **Spatially Varying Gene Regulatory Networks via Bayesian Nonparametric Covariate-Dependent Directed Cyclic Graphical Models** [Abstract 60]

  Trisha DAWN, *Texas A & M University*
  Yang NI,

- **Are ML methods for supervised classification truly better learners than classical linear discriminant analysis?** [Abstract 70]

  Oluwafunmibi FASANYA, *University of Nebraska Lincoln*

- **Bayesian nonparametric common atoms approach for creating synthetic controls in early-phase glioblastoma trials** [Abstract 76]

  Bhanu GARG, *University of Texas at Dallas*
  Noirrit Kiran CHANDRA,
  Lorenzo TRIPPA,
  Peter MÜLLER,
  Rifaquat Musaffa RAHMAN,
  John DE GROOT,

- **PLRD: Partially Linear Regression Discontinuity Inference** [Abstract 79]

  Aditya GHOSH, *Stanford University*
  Aditya GHOSH, *Stanford University*
  Guido IMBENS, *Stanford University*
  Stefan WAGER, *Stanford University*

- **Evaluation of Cox Mixture Models for End-Stage Kidney Disease for Higher Risk Patients** [Abstract 94]

  Jason R HASSE, *South Dakota State University*
  Semhar MICHAEL,

- **Bayesian Sparse Regression for Microbiome-Metabolite Data Integration** [Abstract 108]

  Kai JIANG, *The University of Texas Health Science Center at Houston*
  Satabdi SAHA, *The University of Texas MD Anderson Cancer Center*
  Christine PETERSON, *The University of Texas MD Anderson Cancer Center*

- **Adaptive Divide and Conquer with Two Rounds of Communication** [Abstract 112]

  Niladri KAL, *Texas A&M University*
  Debdeep PATI, *University of Wisconsin-Madison*
  Botond SZABO, *Bocconi University*
  Rajarshi GUHANIYOGI, *Texas A&M University*

- **SMART-MC: Sparse Matrix Estimation with Covariate-Based Transitions in Markov Chain Modeling of Multiple Sclerosis Disease Modifying Therapies** [Abstract 121]

Beomchang KIM, *Virginia Commonwealth University*
Priyam DAS, *Virginia Commonwealth University*
Zongqi XIA, *University of Pittsburgh*

- **Distribution Regression Using Conditional Deep Generative Models** [Abstract 126]

Shivam KUMAR, *University of Notre Dame*
Shivam KUMAR, *University of Notre Dame*
Yun YANG, *University of Maryland at College Park*
Lizhen LIN, *University of Maryland at College Park*

- **Structure Learning and Statistical Inference in Tensor Ising Models** [Abstract 137]

Tianyu LIU, *National University of Singapore*
Tianyu LIU, *Department of Statistics and Data Science, National University of Singapore*
Somabha MUKHERJEE, *Department of Statistics and Data Science, National University of Singapore*
Bhaswar BHATTACHARYA, *Department of Statistics, University of Pennsylvania*

- **Computationally efficient reductions between some statistical models** [Abstract 139]

Mengqi LOU, *Georgia Institute of Technology*
Guy BRESLER, *MIT*
Ashwin PANANJADY, *Georgia Institute of Technology*

- **Bayesian Models for Joint Selection of Features and Auto-Regressive Lags: Theory and Applications in Environmental and Financial Forecasting** [Abstract 152]

Alokesh MANNA, *University of Connecticut*
Sujit GHOSH,

- **Estimating stationary mass, frequency by frequency** [Abstract 170]

Milind NAKUL, *Georgia Institute of Technology*
Vidya MUTHUKUMAR, *Georgia Institute of Technology*
Ashwin PANANJADY, *Georgia Institute of Technology*

- **Bayesian Cooperative Learning for Multimodal Integration** [Abstract 199]

Saptarshi ROY, *Texas A&M University*
Sreya SARKAR, *University of Iowa*
Himel MALLICK, *Cornell University*
Nengjun YI, *University of Alabama, Birmingham*

- **Bayesian interpretation of a second-order efficient empirical Bayes confidence interval** [Abstract 211]

Aditi SEN, *Department of Mathematics, University of Maryland, College Park*
Masayo Y. HIROSE, *Institute of Mathematics for Industry, Kyushu University*
Partha LAHIRI, *Department of Mathematics, University of Maryland, College Park*

- **Optimal Use of Survey Weights for Causal Inference under Informative Sampling** [Abstract 212]

Shubhajit SEN, *North Carolina State University*
Shu YANG, *Professor of Statistics*

Special Invited Session 1 Bo Li, Yongming Qu

Venue: Plains A

Chair : Hiya BANERJEE, Eli Lilly and Company

14:45 Perspectives on Climate Model Evaluation [Abstract 130]

Bo LI, *Washington University in St. Louis*

15:30 Advancing Drug Development with Statistical Innovation - My Perspective [Abstract 191]

Yongming QU, *Eli Lilly and Company*

01.A1.I5 Harnessing AI and Advanced Analytics in Clinical Development: From Insights to Innovation

Venue: Plains B

Chair : Sanhita SENGUPTA, Bristol Myers Squibb

Organizer : Jianchang LIN, Takeda

14:45 AI-powered Clinical Development: Integrating Clinical Insights with Statistical Innovations [Abstract 145]

Will MA, *HopeAI*

15:15 AI-Powered Surrogate Endpoint Validation for Oncology Trials: A Case Study in Multiple Myeloma [Abstract 243]

Lanjing WANG, *University of Washington*
Zixuan ZHAO, *George Washington University*
Zexin REN, *George Washington University*
Will MA, *Hope AI*

15:45 Bayesian Dynamic Borrowing Meets Machine Learning: A Case Study of Hybrid Control Arms in Oncology Trial [Abstract 214]

Sanhita SENGUPTA, *Bristol Myers Squibb*
Jixian WANG, *Bristol Myers Squibb*
Ram TIWARI, *HopeAI*

01.A1.I6 New Advances in Causal Inference

Venue: Plains C

Chair : Chan PARK, University of Illinois Urbana-Champaign

Organizer : Rajen SHAH, University of Cambridge

14:45 Long-term causal inference under persistent confounding via data combination [Abstract 247]

Yuhao WANG, *Tsinghua University*
Guido IMBENS, *Stanford University*
Nathan KALLUS, *Cornell University*
Xiaojie MAO, *Tsinghua University*

15:15 Optimization-based Sensitivity Analysis [Abstract 71]

Tobias FREIDLING, *École polytechnique fédérale de Lausanne*
Qingyuan ZHAO, *University of Cambridge*

15:45 Proximal Causal Inference for Conditional Separable Effects [Abstract 180]

Chan PARK, *University of Illinois Urbana-Champaign*
Mats STENSRUD, *EPFL*
Eric TCHETGEN TCHETGEN, *University of Pennsylvania*

01.A1.I7 High-dimensional data in theory and applications

Venue: Prairie A

Chair and Organizer : Nabarun DEB, University of Chicago

14:45 What does Guidance do in Masked Discrete Diffusion Models [Abstract 95]

Ye HE, *Georgia Institute of Technology*
Kevin ROJAS, *Georgia Institute of Technology*
Molei TAO, *Georgia Institute of Technology*

15:15 TBD [Abstract 162]

Debarghya MUKHERJEE, *Boston University*

15:45 Polyspectral Mean Estimation of General Nonlinear Processes [Abstract 81]

Dhrubajyoti GHOSH, *Duke University*
Tucker MCELROY, *U.S. Census Bureau*
Soumendra LAHIRI, *Washington University in St. Louis*

01.A1.I8 Networks and Graphical Models

Venue: Prairie B

Chair : Sagnik NANDY, University of Chicago
Organizer : Bhaswar BHATTACHARYA, University of Pennsylvania

14:45 Propagation of Shocks on Networks: Can Local Information Predict Survival? [Abstract 65]

Souvik DHARA, *Purdue University*
Manish PANDEY,
Leonard SCHULMAN, *CalTech*

15:15 Conformal Prediction for Dyadic Regression under Structured Missingness [Abstract 142]

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

15:45 Mixing Phases and Metastability of the Glauber Dynamics in Tensor Curie-Weiss Models [Abstract 165]

Somabha MUKHERJEE, *National University of Singapore*
Ramkrishna SAMANTA, *University College London*
Jiang ZHANG, *National University of Singapore*

01.A1.I9 Enhancing Clarity and Decision-Making in Data Science: From Environmental Clustering to AI in Genomics

Venue: Prairie C

Chair : Venkata Sai Pramod Kumar KASTURI, Corteva Agriscience
Organizer : Sagar KSHEERA, Corteva

14:45 How environmental clustering revealed confusion in the statistical literature and how we fixed it [Abstract 248]

Kevin WRIGHT, *Corteva Agriscience*

15:15 Box-Cox Transformation - 60 Years After [Abstract 194]

Marepalli RAO, *University of Cincinnati*
Nisha SHESHASHAYEE, *University of Cincinnati*
Wedad ALATEBI, *University of Cincinnati*
Tianyuan GUAN, *Kent State University*

15:45 **Quantifying Uncertainty in Crop Yield Predictions using Deep Learning Ensembles for Risk-Informed Decision-Making** [Abstract 117]

Venkata Sai Pramod Kumar KASTURI, *Corteva Agriscience*
Bishwa SAPKOTA, *Corteva Agriscience*
Venkat NEMANI, *Corteva Agriscience*
Hoda HELMI, *Corteva Agriscience*

01.A1.I10 Recent Advances in Bayesian Methods

Venue: Arbor

Chair and Organizer : Joyee GHOSH, The University of Iowa

14:45 **Bayesian structured variable selection in finite mixture of regression analysis for cancer data** [Abstract 103]

Yunju IM, *University of Nebraska Medical Center*

15:15 **B-MASTER: Scalable Bayesian multivariate regression analysis for selecting targeted essential regressors to identify the key genera in microbiome-metabolite relation dynamics** [Abstract 56]

Priyam DAS, *Virginia Commonwealth University*

15:45 **TBD** [Abstract 12]

Sanjib BASU, *University of Illinois at Chicago*

01.A1.I11 Advances in Bayesian Spatio-temporal and Extreme Value Modeling

Venue: Garden

Chair : Snigdhansu CHATTERJEE, University of Maryland, Baltimore County
Organizer : Sakshi ARYA, Case Western Reserve University

14:45 **TBD** [Abstract 44]

Snigdhansu CHATTERJEE, *University of Maryland, Baltimore County*

15:15 **Bayesian Hierarchical Modeling for Extremes** [Abstract 7]

Sakshi ARYA, *Case Western Reserve University*
Snigdhanshu CHATTERJEE, *UMBC*
Vishal SUBEDI, *UMBC*
Dietz LINDSEY, *University of Minnesota*

15:45 **Point Prediction of Streaming Data** [Abstract 37]

Aleena CHANDA, *University of Nebraska-Lincoln*
Bertrand CLARKE, *University of Nebraska-Lincoln*

Time : **16:15 - 16:30** **Coffee Break** *Venue:* Foyer

01.E1.I12 Advances in uncertainty quantification in Machine Learning

Venue: Plains A

Chair : Arun KUCHIBHOTLA, Carnegie Mellon University
Organizer : Purnamrita SARKAR, UT Austin

16:30 **Uncertainty Quantification for Functionals in Constrained Inverse Problems** [Abstract 184]

Pratik PATIL, *University of California, Berkeley*

17:00 **Constructing confidence sequences from adaptive Robbins-Siegmund's lemma** [Abstract 238]

Pham TUAN, *UT Austin*

17:30 **Sequential change detection with simulators** [Abstract 217]

Shubhanshu SHEKHAR, *University of Michigan, Ann Arbor*

01.E1.I13 Innovative Statistical and AI Approaches in Public Health and Behavioral Research

Venue: Plains B
Chair : Aiying ZHANG, University of Virginia
Organizer : Li WANG, Abbvie

16:30 **Marginal Structural Modeling for Causal Inference of E-Cigarette Vaping and Smoking Relapse – Six-Wave Longitudinal Study 2013 - 2021** [Abstract 53]

Daisy DAI, *UNMC*

17:00 **Wasserstein Geodesic Generator for Conditional Distributions** [Abstract 122]

Younggeun KIM, *Michigan State University*

17:30 **GAE-BEG Model: A novel GNN Framework integrating neuroimaging and behavioral information to understand Adolescent Psychiatric Disorders** [Abstract 259]

Aiying ZHANG, *University of Virginia*
Gang QU, *Tulane University*

01.E1.I14 Recent advances in observational studies

Venue: Plains C
Chair : Mengxin YU, University of Pennsylvania
Organizer : Tirthankar DASGUPTA, Rutgers University

16:30 **Proximal Causal Inference with Some Invalid Proxies** [Abstract 193]

Prabrisha RAKSHIT, *University of Pennsylvania*
Eric TCHETGEN TCHETGEN, *University of Pennsylvania*
Xu SHI, *University of Michigan*

17:00 **Test-negative designs with various reasons for testing: statistical bias and solution** [Abstract 256]

Mengxin YU, *University of Pennsylvania*

01.E1.I15 Recent advances in high-dimensional statistics and inference

Venue: Prairie A
Chair : Surya TOKDAR, Duke University
Organizer : Miaoyan WANG, University of Wisconsin - Madison

16:30 **Inference with Randomized Regression Trees** [Abstract 179]

Snigdha PANIGRAHI, *University of Michigan*
Soham BAKSHI, *University of Michigan*
Yiling HUANG, *University of Michigan*
Walter DEMPSEY, *University of Michigan*

17:00 **Nonparametric Empirical Bayes and Selective Inference** [Abstract 235]

Surya TOKDAR, *Duke University*
Peter HOFF, *Duke University*

01.E1.I16 Processes on Networks

Venue: Prairie B
Chair and Organizer : Souvik DHARA, Purdue University

16:30 **The Effect of Restrictive Interactions between Susceptible and Infected Individuals on the Prognosis of an Epidemic** [Abstract 43]

Shirshendu CHATTERJEE, *City University of New York*

17:00 **Long-range competition on the torus** [Abstract 148]

Neeladri MAITRA, *University of Illinois at Urbana-Champaign*
Bas LODEWIJKS,

17:30 **Large Deviation Results for Time Averages of a Rapidly Evolving Dynamic W-Random Graph Model** [Abstract 195]

Souvik RAY, *School of Data Science & Society, University of North Carolina at Chapel Hill*
Amarjit BUDHIRAJA, *University of North Carolina at Chapel Hill*
Shankar BHAMIDI, *University of North Carolina at Chapel Hill*

01.E1.I17 Recent advances in high-dimensional learning

Venue: Prairie C

Chair and Organizer : Subhabrata SEN, Harvard University

16:30 **Universality Phenomenon in Random Feature and Kernel-based Learning** [Abstract 101]

Hong HU, *Washington University in St. Louis*

17:00 **High-dimensional Asymptotics of Differentially Private PCA** [Abstract 68]

Rishabh DUDEJA, *UW Madison*
Youngjoo YUN, *UW Madison*

17:30 **Universality of Max-Margin Classifiers** [Abstract 223]

Youngtak SOHN, *Brown University*
Andrea MONTANARI, *Stanford University*
Feng RUAN, *Northwestern University*
Basil SAEED, *Stanford University*

01.E1.I18 Approximate algorithms for complex Bayesian problems

Venue: Arbor

Chair : Prateek JAISWAL, Purdue University
Organizer : Debdeep PATI, University of Wisconsin-Madison

16:30 **CLT in HD Bayesian linear regression** [Abstract 166]

Sumit MUKHERJEE, *Columbia University*
Seunghyun LEE, *Columbia University*
Nabarun DEB, *University of Chicago*

17:00 **Preference Optimization on Pareto Sets: On a Theory of Multi-Objective Optimization** [Abstract 198]

Abhishek ROY, *Texas A&M University*
Geelon SO, *University of California San Diego*
Yi-An MA, *University of California San Diego*

17:30 **Generalizing Regret bounds for Thompson sampling through Minimax-Optimal α-Posterior Concentration Analysis** [Abstract 104]

Prateek JAISWAL, *Purdue University*
Debdeep PATI, *Department of Statistics, University of Wisconsin-Madison*
Anirban BHATTACHARYA, *Department of Statistics, Texas A&M University*
Bani MALLICK, *Department of Statistics, Texas A&M University*

01.E1.I19 Design and Optimization

Venue: Garden

Chair : Evan ROSENMAN, Claremont McKenna College
Organizer : Abhyuday MANDAL, University of Georgia

16:30 **F-modeling-Based empirical Bayes Estimator for Parameters in the Scale Family** [**Abstract 127**]

YEIL KWON, *Wichita State University*

17:00 **Adaptive Experimental Design Using Shrinkage Estimators** [**Abstract 197**]

Evan ROSENMAN, *Claremont McKenna College*
Kristen HUNTER, *University of New South Wales*

Friday June 13

Short Course 1 Boosting R Code performance via C++ Integration within Rstudio *Venue:* Garden

Time : **9:00 - 12:30**

- **Priyam DAS**, Virginia Commonwealth University

02.M1.I20 Dr. Riten Mitra Memorial Session *Venue:* Plains A

Time : **9:00 - 10:30**

Chair and Organizer : Subhajit SENGUPTA, Cytel Inc.

- **Model-based inference for multiple dependent graphs – Riten Mitra's contributions** [**Abstract 161**]

  Peter MUELLER, *Professor, Department of Statistics and Data Sciences, Department of Mathematics, UT Austin*

- **Bayesian Nonparametric Methods for Oncology Studies: Riten Mitra's Impact in Cancer Research** [**Abstract 251**]

  Yanxun XU, *Associate Professor, Department of applied mathematics and statistics, Johns Hopkins University*

- **Statistical Learning of SDEs: A Journey with Riten Mitra** [**Abstract 75**]

  Arnab GANGULY, *Associate Professor, Department of Mathematics, Louisiana State University*

- **"Riten Da : sometimes, it's ok!"** [**Abstract 17**]

  Sagnik BHADURY, *University of Michigan*

02.M1.I21 Statistical Methods for Networks, Tensors, and Bedyond *Venue:* Plains B

Chair : Joshua AGTERBERG, University of Illinois Urbana-Champaign

Organizer : Miaoyan WANG, University of Wisconsin - Madison

9:00 **Tensor-on-Tensor Times Series Regression for Integrated One-step Analysis of fMRI Data** [**Abstract 149**]

Ranjan MAITRA, *Iowa State University*
Subrata PAL,

9:30 **Network two-sample test for block models** [**Abstract 147**]

Oscar Hernan MADRID PADILLA, *University of California, Los Angeles*
Chung Kyong NGUEN, *University of California, Los Angeles*
Arash AMINI,

10:00 **Estimation and Inference in Tensor Mixed-Membership Blockmodels** [**Abstract 3**]

Joshua AGTERBERG, *University of Illinois Urbana-Champaign*
Anru ZHANG, *Duke University*

02.M1.I22 Recent advances in causal inference *Venue:* Plains C

Chair and Organizer : Subhabrata SEN, Harvard University

9:00 **Estimating the Global Average Treatment Effect under Structured Interference** [**Abstract 133**]

Shuangning LI, *University of Chicago*
Kevin HAN, *Meta*
Johan UGANDER, *Stanford University*

9:30 **The Decaying MAR Framework: Model Doubly Robust Causal Inference with Partially Labeled Data** [Abstract 29]

Abhishek CHAKRABORTTY, *Texas A&M University*

10:00 **Agnostic Characterization of Interference in Randomized Experiments** [Abstract 50]

David CHOI, *Carnegie Mellon University*

02.M1.I23 Advances in Bias Correction, Robust Clustering, and Indirect Comparisons in Observational Studies

Venue: Prairie A

Chair and Organizer : Anirban MONDAL, Case Western Reserve University

9:00 **Covariate adjustment for marginal estimates in randomized trials: a primer and extension to interval-censored outcomes.** [Abstract 221]

Richard SIZELOVE, *Eli Lilly & Company*

9:30 **Correcting Latent Class Confounder Bias in Observational Studies** [Abstract 222]

Abdul-Nasah SOALE, *Case Western Reserve University*

10:00 **Robust K-means Clustering based on Density Power Divergence Measure** [Abstract 10]

Paromita BANERJEE, *John Carroll University*
Anirban MONDAL, *Case Western Reserve University*
Abhijit MANDAL, *University of Texas at El Paso*

02.M1.C1 High dimensional data in Theory and Applications

Venue: Prairie B

Chair and Organizer : Nabarun DEB, University of Chicago

9:00 **Inference in Federated Learning: limit theorems and applications** [Abstract 22]

Soham BONNERJEE, *University of Chicago*
Sayar KARMAKAR, *University of Florida*

9:15 **Making prediction intervals smarter: randomization enables local coverage** [Abstract 99]

Rohan HORE, *University of Chicago*
Rina BARBER, *University of Chicago*

9:30 **Concentration inequalities for correlated network-valued processes with applications to community estimation and changepoint analysis** [Abstract 172]

Anirban NATH, *Columbia University*

9:45 **Expectation Maximization Estimation for Hawkes Process with Missingness** [Abstract 255]

Jingtian YU, *Oregon State University*
Jingtian YU, *Oregon State university*
Sharmodeep BHATTACHARYYA, *Oregon State university*
Sarah EMERSON, *Oregon State university*

10:00 **Frequency Domain Resampling for Gridded Spatial Data** [Abstract 15]

Souvick BERA, *Colorado School of Mines*
Daniel J. NORDMAN, *Iowa State University*
Soutir BANDYOPADHYAY, *Colorado School of Mines*

02.M1.I24 Statistics and Generative AI

Venue: Prairie C

Chair and Organizer : Subho MAJUMDAR, Vijil

9:00 **HEART: Heterogeneous data-driven Emotion and Anomaly Recognition in Sparse Longitudinal Texts** [Abstract 87]

Aritra GUHA, *AT&T Chief Data Office*
Prasanjit DUBEY, *Georgia Institute of Technology*
Paromita DUBEY, *University of Southern California*
Zhengyi ZHOU, *AT&T*

9:30 **Measuring and Ensuring Consistency in Language Models** [Abstract 150]

Subho MAJUMDAR, *Vijil*
Harsh RAJ, *Northeastern University*
Domenic ROSATI, *Dalhousie University*
Vipul GUPTA, *Penn State University*

02.M1.I25 Modern Bayesian methods for public health

Venue: Arbor

Chair : Arkajyoti SAHA, University of California, Irvine
Organizer : Abhi DATTA, Johns Hopkins University

9:00 **Tree-Regularized Bayesian Latent Class Analysis for Improving Weakly Separated Dietary Pattern Subtyping in Small-Sized Subpopulations** [Abstract 250]

Zhenke WU, *Department of Biostatistics, University of Michigan*
Mengbing LI, *University of Michigan*
Briana STEPHENSON, *Harvard University*

9:30 **Bayesian Modeling of Misclassification Matrices forImproved Verbal Autopsy-Based Mortality Estimatesin LMICs** [Abstract 189]

Sandipan PRAMANIK, *Johns Hopkins Bloomberg School of Public Health*
Scott ZEGER, *Johns Hopkins Bloomberg School of Public Health*
Dianna BLAU, *Global Health Center, US Centers for Disease Control and Prevention*
Abhirup DATTA, *Johns Hopkins Bloomberg School of Public Health*

10:00 **Bayesian Federated Cause-of-Death Quantification Under Distribution Shift** [Abstract 257]

Li ZEHANG, *University of California, Santa Cruz*
Yu ZHU, *University of California, Santa Cruz*

Time : **10:30 - 10:45** **Coffee Break** *Venue:* Foyer

Special Invited Session 2 Adityanand Guntuboyina, Kengo Kato

Venue: Plains A

Chair : Bodhisattva SEN, Columbia University

10:45 **Totally concave regression** [Abstract 90]

Adityanand GUNTUBOYINA, *University of California Berkeley*
Dohyeong KI, *University of California Berkeley*

11:30 **Inference with Gromov-Wasserstein Distances** [Abstract 118]

Kengo KATO, *Cornell University*
Gabriel RIOUX, *Cornell University*
Ziv GOLDFELD, *Cornell University*

Panel Discussion 2 The Future of Statistics in This New World of AI

Venue: Plains B

Time : **10:45 - 12:15**

Moderator: Hiya BANERJEE, Eli Lilly and Company

- **Martin FRENZEL**, Eli Lilly
- **Subho MAJUMDAR**, Vijil
- **Tim FRIEDE**, University Medical Center Göttingen
- **Kris SANKARAN**, University of Wisconsin-Madison
- **Jason KLUSOWSKI**, Princeton University

02.M2.I26 Assumption-lean Inference

Venue: Plains C

Chair : Nilanjana LAHA, Texas A&M

Organizer : Rajen SHAH, University of Cambridge

10:45 Higher-order estimators of time-varying effects in anisotropic smoothness models [Abstract 23]

Matteo BONVINI, *Rutgers University*
Edward KENNEDY, *Carnegie Mellon University*
Luke KEELE, *University of Pennsylvania*

11:15 Nonparametric inference on non-negative dissimilarity measures at the boundary of the parameter space [Abstract 102]

Aaron HUDSON, *Fred Hutchinson Cancer Center*

11:45 Model-free dynamic treatment regimes with arbitrary number of treatments and stages [Abstract 128]

Nilanjana LAHA, *Texas A&M*
Nilson CHAPAGAIN, *Texas A&M*
Victoria CICHERSKI, *HEB*
Aaron SONABEND-W, *Google Research*

02.M2.I27 AI-Driven Innovations in Statistical Analysis and Clinical Trial Design

Venue: Prairie A

Chair : Tusharkanti GHOSH, Colorado School of Public Health

Organizer : Abishek BHATTACHARJEE, Pfizer

10:45 AI-powered Flexible-design Simulation Assistant for Clinical Trials [Abstract 215]

Subhajit SENGUPTA, *Cytel Inc.*
Kyle WATHEN, *Cytel Inc.*
Gabriel POTVIN, *Cytel Inc.*
Anoop Singh RAWAT, *Cytel Inc.*

11:15 A Robust Kernel Machine Framework for Assessing Spatial Variability and Cross-Niche Communication in Spatial Transcriptomics [Abstract 85]

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

02.M2.I28 Statistics in Bioinformatics and Genetics

Venue: Prairie B

Chair and Organizer : Somak DUTTA, Iowa State University

10:45 A Branching Process Model for Digital Read Quantification with Application to PCR-Based Diagnostics [Abstract 66]

Karin S DORMAN, *Iowa State University*
Debosmita KUNDU, *Iowa State University*

11:15 Genetic fine mapping of high-dimensional traits, with application to metabolite genome-wide association studies [Abstract 154]

Chris MCKENNAN, *University of Pittsburgh*
Weiqiong HUANG, *University of Pittsburgh*
Emily HECTOR, *North Carolina State University*

11:45 Scalable divergence time estimation via Hamiltonian Monte Carlo sampling [Abstract 107]

Xiang JI, *Tulane University*

02.M2.I29 Iterative methods in statistical machine learning

Venue: Prairie C

Chair and Organizer : Ashwin PANANJADY, Georgia Tech

10:45 Recent theoretical advances in diffusion models [Abstract 249]

Yuchen WU, *University of Pennsylvania*
Andrea MONTANARI, *Stanford University*
Yuxin CHEN, *University of Pennsylvania*
Yuting WEI, *University of Pennsylvania*

11:15 Minimax Optimality of the Probability Flow ODE for Diffusion Models [Abstract 26]

Changxiao CAI, *University of Michigan*
Gen LI, *The Chinese University of Hong Kong*

11:45 State evolution beyond first order methods [Abstract 178]

Ashwin PANANJADY, *Georgia Tech*
Michael CELENTANO, *OpenAI*
Chen CHENG, *Stanford University*
Kabir VERCHAND, *Georgia Tech*

02.M2.I30 Bayesian Structure Learning with Dependent Data

Venue: Arbor

Chair : Zhang SHUANGJIE, Texas A & M University
Organizer : Anirban BHATTACHARYA, Texas A&M University

10:45 Consistent DAG selection for Bayesian causal discovery under general error distributions [Abstract 48]

Anamitra CHAUDHURI, *Texas A&M University*
Anirban BHATTACHARYA, *Texas A&M University*
Yang NI, *Texas A&M University*

11:15 TBD [Abstract 32]

Antik CHAKRABORTY, *Purdue University*

11:45 Sparse Bayesian Group Factor Model for Feature Interactions in Multiple Count Tables Data [Abstract 220]

Zhang SHUANGJIE, *Texas A & M University*
Shen YUNING, *Department of Chemical and Biomolecular Engineering, UCLA*

Chen IRENE A., *Department of Chemical and Biomolecular Engineering, UCLA*
Lee JUHEE, *Department of Statistics, UCSC*

Time : **12:15 - 13:30 Lunch** *Venue:* **East Campus Dining Center**

Plenary Lecture 2 Linda Young

Venue: Plains A

Chair : Bertrand CLARKE, University of Nebraska-Lincoln

13:30 The Evolution of Crops County Estimates: Past, Present and Future [Abstract 254]

Linda YOUNG,

14:30 - 14:45 Short Break No Coffee

Special Invited Session 3 Susmita Datta, Tim Friede

Venue: Plains A

Chair : Saonli BASU, University of Minnesota

14:45 "Unraveling Disease Mysteries: Statistical Models Reveal Cellular Conversations using Spatial Transcriptomics data." [Abstract 59]

Susmita DATTA, *University of Florida*
Dongyuan WU, *Moderna Inc. and University of Florida*

15:30 (Bayesian) meta-analysis: statistical methods and their applications in clinical medicine [Abstract 72]

Tim FRIEDE, *University Medical Center Göttingen*

02.A1.I31 Innovative Statistical and AI-Driven Approaches for Complex Data Modeling

Venue: Plains B

Chair : Semhar MICHAEL, South Dakota State University
Organizer : Himel MALLICK, Cornell University

14:45 Mixed Poisson families with real-valued mixing distributions [Abstract 237]

Will TOWNES, *Carnegie Mellon University*

15:15 Large scale few-shot learning through parameter pooling [Abstract 156]

Semhar MICHAEL, *South Dakota State University*
Andrew SIMPSON, *South Dakota State University*

02.A1.I32 Causal inference in randomized experiments

Venue: Plains C

Chair : Bingkai WANG, Department of Biostatistics, University of Michigan
Organizer : Peng DING, University of California Berkeley

14:45 Neymanian inference in randomized experiments [Abstract 46]

Ambarish CHATTOPADHYAY, *Stanford University*
Guido IMBENS, *Stanford University*

15:15 Cluster-robust inference with a single treated cluster [Abstract 134]

Xinran LI, *University of Chicago*
Chun Pong LAU, *University of Chicago*
Xinran LI, *University of Chicago*

15:45 TBD [Abstract 239]

Bingkai WANG, *Department of Biostatistics, University of Michigan*

02.A1.I33 Recent Advancements in Spatial Statistics

Venue: Prairie A

Chair and Organizer : Indranil SAHOO, Virginia Commonwealth University

14:45 LatticeVision: Image-to-Image Networks for Modeling Non-Stationary Spatial Data [Abstract 155]

Daniel MCKENZIE, *Colorado School of Mines*
Antony SIKORSKI, *Colorado School of Mines*
Michael IVANITSKIY, *Colorado School of Mines*
Doug NYCHKA, *Colorado School of Mines*

15:15 Dimension Reduction for Regression Model with Spatially Correlated Data [Abstract 160]

Hossein MORADI REKABDARKOLAEE, *South Dakota State University*

15:45 Modeling Spatio-Temporal Environmental Processes and Emerging Crises under Data Sparsity [Abstract 5]

Ali ARAB, *Georgetown University, Department of Mathematics and Statistics*

02.A1.I34 Emerging perspectives in Statistical Learning

Venue: Prairie B

Chair and Organizer : Adityanand GUNTUBOYINA, University of California Berkeley

14:45 Revisiting Total Variation Denoising: New Perspectives and Generalizations [Abstract 41]

Sabyasachi CHATTERJEE, *University of Illinois at Urbana Champaign*

15:15 Can Empirical Bayes via Empirical Risk Minimization Balance Computational and Theoretical Benefits? [Abstract 106]

Soham JANA, *University of Notre Dame*
Yury POLYANSKIY, *MIT*
Anzo TEH, *MIT*
Yihong WU, *Yale University*

15:45 Confidence Interval Construction and Conditional Variance Estimation with Dense ReLU Networks [Abstract 146]

Carlos Misael MADRID PADILLA, *Washington University in St Louis*
Oscar Hernan MADRID PADILLA,
Yik Lun KEI,
Yanzhen CHEN,

02.A1.I35 Advances in high dimensional statistical learning

Venue: Prairie C

Chair and Organizer : Ashwin PANANJADY, Georgia Tech

14:45 Instance-optimal stochastic optimization: Succeeding when sample average approximation fails [Abstract 109]

Liwei JIANG, *Georgia Institute of Technology*

15:15 Distribution-free Inference for Model Class Risk [Abstract 143]

Yuetian LUO, *University of Chicago*
Manuel MULLER, *University of Cambridge*
Rina FOYGEL BARBER, *University of Chicago*

15:45 TBD [Abstract 169]

Vidya MUTHUKUMAR, *Georgia Tech*

02.A1.I36 Innovative Bayesian Approaches for Data Integration and Predictive Modeling

Venue: Arbor
Chair : Subharup GUHA, University of Florida
Organizer : Sagar KSHEERA, Corteva

14:45 Refining Clinical Trials: Calibrated Bayesian Extrapolation Using Data-Informed Priors [Abstract 125]

Maria KUDELA, *Pfizer*
Heliang SHI, *Pfizer*
Yuxi ZHAO, *Pfizer*
Margaret GAMALO, *Pfizer*

15:15 Bayesian Estimation of Propensity Scores for Integrating Multiple Cohorts with High-Dimensional Covariates [Abstract 89]

Subharup GUHA, *University of Florida*
Subharup GUHA, *University of Florida*
Yi LI, *University of Michigan*

02.A1.I37 Methods for Big, Complex Biological Data

Venue: Garden
Chair and Organizer : Karin S DORMAN, Iowa State University

14:45 A hybrid mixture of factor analyzers approach for characterizing high dimensional data [Abstract 54]

Fan DAI, *Michigan Technological University*
Kazeem KAREEM, *Michigan Technological University*

15:15 Quantifying the Mediation Effect for Non-sparse High-dimensional Omics Mediators [Abstract 131]

Chunlin LI, *Iowa State University*

15:45 Decoding Spatial Tissue Architecture: A Scalable Bayesian Topic Model for Multiplexed Imaging Analysis [Abstract 186]

Xiyu PENG, *Texas A&M University*
James SMITHY, *Memorial Sloan Kettering Cancer Center*
Katherine PANAGEAS, *Memorial Sloan Kettering Cancer Center*
Ronglai SHEN, *Memorial Sloan Kettering Cancer Center*

Time : **16:15 - 16:30 Coffee Break** *Venue:* Foyer

02.E1.I38 Advances in Learning & Inference with Complex Data: Networks, Functional Data, and Beyond

Venue: Plains A
Chair and Organizer : Paromita DUBEY, University of Southern California

16:30 Network-linked high-dimensional multinomial Probit [Abstract 163]

Gourab MUKHERJEE, *University of Southern California*
Rashmi BHUYIAN, *USC*
Adel JAVANMARD, *USC*

17:00 Efficient Analysis of Latent Spaces in Heterogeneous Networks [Abstract 96]

Yinqiu HE, *University of Wisconsin-Madison*
Tian YUANG, *Fudan University*
Jiajin SUN, *Florida State University*

17:30 **Testing independence for sparse longitudinal data** [Abstract 266]

Changbo ZHU, *University of Notre Dame*
Wang JANE-LING,

02.E1.I39 Advancing Evidence Generation: Methods for External Control, Causal Inference, and Dynamic Borrowing in Clinical Research

Venue: Plains B

Chair : Soumik PURKAYASTHA, University of Pittsburgh
Organizer : Margaret GAMALO, Pfizer

16:30 **The Role of Propensity Score in Leveraging External Data for Regulatory Decision-Making** [Abstract 240]

CG WANG, *Regeneron*

17:00 **Bayesian randomized basket trial design: a case study from the ultra-rare invasive mould infections** [Abstract 201]

Satrajit ROYCHOUDHURY, *Pfizer Inc.*

17:30 **Examining Directional Association between Depression and Anxiety** [Abstract 190]

Soumik PURKAYASTHA, *University of Pittsburgh*
Peter X.-K. SONG, *University of Michigan*

02.E1.I40 Frontiers in Adaptive Statistical Inference

Venue: Plains C

Chair and Organizer : Adityanand GUNTUBOYINA, University of California Berkeley

16:30 **Adaptive Inference Techniques for Some Irregular Problems** [Abstract 124]

Arun KUCHIBHOTLA, *Carnegie Mellon University*
Woonyoung CHANG, *Carnegie Mellon University*

17:00 **Revisiting empirical risk minimization: new risk characterizations and suboptimality results** [Abstract 183]

Reese PATHAK, *UC Berkeley*

02.E1.I41 Kernel Methods for Nonparametric Inference

Venue: Prairie A

Chair : Somabha MUKHERJEE, National University of Singapore
Organizer : Bhaswar BHATTACHARYA, University of Pennsylvania

16:30 **EXTREMAL EIGENVALUES OF RANDOM KERNEL MATRICES WITH POLYNOMIAL SCALING** [Abstract 171]

Sagnik NANDY, *University of Chicago*
David KOGAN, *Yale University*
Jiaoyang HUANG, *University of Pennsylvania*

17:00 **Computational-Statistical Trade-offs in Kernel Two-Sample Testing** [Abstract 226]

Bharath SRIPERUMBUDUR, *Pennsylvania State University*
Soumya MUKHERJEE, *Pennsylvania State University*

17:30 **Distance and Kernel-Based Measures for Global and Local Two-Sample Conditional Distribution Testing** [Abstract 261]

Xianyang ZHANG, *Texas A&M University*
Jian YAN, *Cornell University*
Zhuoxi LI, *Xiamen University*

02.E1.I42 Networks: Learning and Inference

Venue: Prairie B

Chair and Organizer : Souvik DHARA, Purdue University

16:30 **Spectral algorithms for community detection in multiview networks** [Abstract 213]

Subhabrata SEN, *Harvard University*
Xiaodong YANG, *Harvard University*
Yue M. LU, *Harvard University*

17:00 **A Poincaré Inequality and Consistency Results for Signal Sampling on Large Graphs** [Abstract 202]

Luana RUIZ, *Johns Hopkins University*
Thien LE,
Stefanie JEGELKA,

02.E1.I43 Semi-parametric and high-dimensional statistics

Venue: Prairie C

Chair : Wooseok HA, KAIST

Organizer : Garvesh RASKUTTI, University of Wisconsin-Madison

16:30 **Semi-Parametric Batched Global Multi-Armed Bandits with Covariates** [Abstract 224]

Hyebin SONG, *Pennsylvania State University*
Sakshi ARYA, *Case Western Reserve University*

17:00 **Non-parametric mixture models for covariance function estimation** [Abstract 16]

Stephen BERG, *Penn State Statistics*
Hyebin SONG,

17:30 **When few labeled target data suffices: a theory of semi-supervised domain adaptation via fine-tuning from multiple starts** [Abstract 92]

Wooseok HA, *KAIST*
Yuansi CHEN, *ETH Zurich*

02.E1.I44 Bayesian Methods and Machine Learning for Dynamic Data Analysis and Prediction

Venue: Arbor

Chair : Pragya SUR, Harvard University

Organizer : Himel MALLICK, Cornell University

16:30 **Online Bayesian Variable Selection for Logistic Regression Models With Streaming Data** [Abstract 82]

Joyee GHOSH, *The University of Iowa*
Shamriddha DE, *The University of Iowa*
Payel GHOSAL, *University of Wisconsin–Madison*

17:00 **Quantifying the effects of transfer learning in min-norm interpolation** [Abstract 229]

Pragya SUR, *Harvard University*
Yanke SONG, *Harvard University*
Sohom BHATTACHARYA, *University of Florida*

02.E1.I45 Recent advances in Spatial Statistics

Venue: Garden

Chair : Sandipan PRAMANIK, Johns Hopkins Bloomberg School of Public Health

Organizer : Abhi DATTA, Johns Hopkins University

16:30 **Matrix-free Conditional Simulation of Gaussian random fields** [Abstract 69]

Somak DUTTA, *Iowa State University*
Debashis MONDAL, *Washington University, St. Louis, MO, USA*

17:00 **Fast and Accurate Fourier Analysis from Irregularly Sampled Data** [**Abstract 78**]

Christopher GEOGA, *University of Wisconsin-Madison*
Paul BECKMAN,

17:30 **Random forests for binary geospatial data** [**Abstract 205**]

Arkajyoti SAHA, *University of California, Irvine*
Abhirup DATTA, *Johns Hopkins Bloomberg School of Public health*

Time : **19:00-22:00** **Banquet Dinner and Award Ceremony** *Venue:* The Graduate by Hilton Hotel

Saturday June 14

Short Course 2 Multimodal Causal Inference for Data Science and Biomedical Research
Venue: Garden
Time : **9:00 - 12:30**

- **Himel MALLICK**, Cornell University

03.M1.I46 Machine Learning in Spatial Extremes: Bridging Spatial Statistics and Extreme Value Theory *Venue:* Plains A
Chair and Organizer : Debjoy THAKUR, Washington University in St. Louis

9:00 **Prediction of Tropical Pacific Rain Rates with Overparameterized Neural Networks** [**Abstract** 111]

Mikyoung JUN, *University of Houston*
Hojun YOU,
Jiayi WANG,
Raymond WONG,

9:30 **Modeling Spatial Extremes using Non-Gaussian Spatial Autoregressive Models via Convolutional Neural Networks** [**Abstract** 9]

Soutir BANDYOPADHYAY, *Colorado School of Mines*
Sweta RAI, *Colorado School of Mines*
Douglas NYCHKA, *Colorado School of Mines*

10:00 **Variable Selection in Spatial Regression: Local LASSO** [**Abstract** 233]

Debjoy THAKUR, *Washington University in St. Louis*
Nwakanma SIDNEY, *Washington University in St. Louis*
Soumendra LAHIRI, *Washington University in St. Louis*

03.M1.I47 Topics in Data Science *Venue:* Plains B
Chair and Organizer : Subhashis GHOSHAL, North Carolina State University

9:00 **Data Science Problems in the Tech Industry** [**Abstract** 168]

Jami MULGRAVE, *North Carolina State University*

9:30 **Large language models in sports analytics** [**Abstract** 218]

Weining SHEN, *University of California, Irvine*

10:00 **Oracle optimal unsupervised Bayesian image segmentation** [**Abstract** 86]

Subhashis GHOSHAL, *North Carolina State University*
Eduard BELITSER, *VU Amsterdam*
Shuvrarghya GHOSH, *North Carolina State University*

03.M1.I48 Recent developments in small area and related topics *Venue:* Plains C
Chair : Jairo Alberto FUQUENE PATINO, Department of Statistics, UC Davis
Organizer : Snigdhansu CHATTERJEE, University of Maryland, Baltimore County

9:00 **A Pseudo-likelihood Approach to Under-5 Mortality Estimation** [**Abstract** 175]

Taylor OKONEK, *Macalester College*
Katherine WILSON, *University of Washington*
Jon WAKEFIELD, *University of Washington*

9:30 **Bayesian Mixture Models, Non-local Prior Formulations and MCMC Algorithms** [**Abstract 73**]

Jairo Alberto FUQUENE PATINO, *Department of Statistics, UC Davis*
Mark STEEL, *University of Warwick*
David ROSSELL, *Universitat Pompeu Fabra in Barcelona*

03.M1.I49 Innovative Bayesian Paradigms: Navigating Optimal Testing, Dynamic Networks, & Beyond

Venue: Prairie A

Chair and Organizer : Partha SARKAR, Florida State University

9:00 **Optimal Estimation and Testing under Horseshoeplus Priors** [**Abstract 83**]

Malay GHOSH, *University of Florida*
Zikun QIN,

9:30 **Generalized Bayesian Inference for Dynamic Random Dot Product Graphs** [**Abstract 140**]

Joshua LOYAL, *Florida State University*
Joshua LOYAL, *Florida State University*

10:00 **Leveraging the blessing of dimensions for scalable Bayesian inference on covariance matrices** [**Abstract 47**]

Shounak CHATTOPADHYAY, *University of California, Los Angeles*
Anru ZHANG, *Duke University*
David DUNSON, *Duke University*

03.M1.I50 Advancing Clinical Trial Design: Bayesian Approaches for Efficiency and Adaptability

Venue: Prairie B

Chair and Organizer : Zhaohua LU, Daiichi-Sankyo Inc.

9:00 **A Bayesian Hybrid Phase 2 Design Incorporating Historical Monotherapy Data and Covariate Adjustment** [**Abstract 141**]

Zhaohua LU, *Daiichi-Sankyo Inc.*
Yiyuan HUANG, *University of Michigan*
Philip HE, *Daiichi-Sankyo*

9:30 **Optimizing Combination Therapies Using a Bayesian Adaptive Design with a Two-dimensional NDLM** [**Abstract 74**]

Byron GAJEWSKI, *University of Kansas Medical Center*

10:00 **Bayesian Optimal Phase II Randomized Clinical Trial Design for Immunotherapy with Delayed Outcomes** [**Abstract 177**]

Haitao PAN, *St Jude*

03.M1.I51 Trustworthy probabilistic inference

Venue: Prairie C

Chair : Abhisek CHAKRABORTY, Eli Lilly and Company
Organizer : Debdeep PATI, University of Wisconsin-Madison

9:00 **Quasi-Bayes in Conditional Moment Restriction Models** [**Abstract 114**]

Sid KANKANALA, *University of Chicago*

9:30 **A Tangent Approximation approach to Variational Inference in Strongly super-Gaussian likelihood models** [**Abstract 64**]

Pritam DEY, *Texas A&M University*
Somjit ROY, *Texas A&M University*
Debdeep PATI, *University of Wisconsin Madison*
Bani MALLICK, *Texas A&M University*

10:00 **Robust probabilistic inference via a constrained transport metric** [**Abstract 31**]

Abhisek CHAKRABORTY, *Eli Lilly and Company*
Abhisek CHAKRABORTY, *Eli Lilly and Company*
Anirban BHATTACHARYA, *Texas A&M University*
Debdeep PATI, *University of Wisconsin, Madison*

03.M1.I52 Bayesian Approaches and Statistical Learning for Complex Data Analysis *Venue:* Arbor

Chair : Siyuan MA, Vanderbilt University Medical Center
Organizer : Himel MALLICK, Cornell University

9:00 **High dimensional mediation analysis with applications in genetics** [**Abstract 260**]

Qi ZHANG, *University of New Hampshire*
Qi ZHANG, *University of New Hampshire*

9:30 **Network models for spatial transcriptomics data** [**Abstract 2**]

Satwik ACHARYYA, *University of Alabama at Birmingham*

10:00 **MiLC for Adjusting of Compositionality and Unobserved Confounding in Microbiome Data** [**Abstract 144**]

Siyuan MA, *Vanderbilt University Medical Center*
Chih-Ting YANG, *Vanderbilt University Medical Center*
Yu SHYR, *Vanderbilt University Medical Center*
Chris MCKENNAN, *University of Pittsburgh*

Time : **10:30 - 10:45** **Coffee Break** *Venue:* **Foyer**

Special Invited Session 4 Dan Nettleton, Jingyi Jessica Li *Venue:* Plains A

Chair : Indranil MUKHOPADHYAY, University of Nebraska-Lincoln

10:45 **Who Is Winning? Determining Whether a Candidate Leads in a Ranked-Choice Election** [**Abstract 173**]

Dan NETTLETON, *Iowa State University*
Shigeki KANAMORI, *Iowa State University*

11:30 **SyNPar: Synthetic Null Parallelism for High-Power and Fast FDR Control in Feature Selection** [**Abstract 132**]

Jingyi Jessica LI, *University of California, Los Angeles*
Changhu WANG, *University of California, Los Angeles*
Ziheng ZHANG, *University of California, Los Angeles*

Panel Discussion 3 Modern Teaching and Career Development in Studying Statistics *Venue:* Plains B

Time : **10:45 - 12:15**
Moderator: Srijata SAMANTA, Bristol Myers Squibb

- **Sanjib BASU**, University of Illinois at Chicago
- **Carmen TEKWE**, Indian University Bloominton
- **Bhaskar BHATTACHARYA**, University of Nebraska-Lincoln
- **Hiya BANERJEE**, Eli Lilly and Company
- **Tim FRIEDE**, University Medical Center Göttingen

03.M2.I53 Advances in Change Point Detection and Time-Dependent Processes

Venue: Plains C

Chair : Mina KARZAND, UC Davis
Organizer : Hao CHEN, University of California, Davis

10:45 High-dimensional Change-point Detection Using Generalized Homogeneity Metrics [Abstract 244]

Runmin WANG, *Texas A&M University*
Shubhadeep CHAKRABORTY, *Bristol Myers Squibb Company*
Xianyang ZHANG, *Texas A&M University*

11:15 Micro-macro changepoint inference for periodic data sequences [Abstract 120]

Rebecca KILLICK, *Lancaster University / UC Santa Cruz*

11:45 Time-stamped Networks and Hawkes Process with Missing Data [Abstract 21]

Sharmodeep BHATTACHARYYA, *Oregon State University*
Sarah EMERSON, *Oregon State University*
Robert TRANGUCCI, *Oregon State University*
Jingtian YU, *Oregon State University*

03.M2.I54 Network resampling and beyond

Venue: Prairie A

Chair : Jonathan STEWART, Florida State University
Organizer : Subrata KUNDU, George Washington University

10:45 Statistical Inference for Subgraph Densities Under Random Sampling from Network Data [Abstract 33]

Nilanjan CHAKRABORTY, *Missouri University of Science and Technology*
Ayoushman BHATTACHARYA, *Washington University in Saint Louis*
Soumen LAHIRI, *Washington University in Saint Louis*

11:15 Network Bootstrap Using Overlapping Partitions [Abstract 28]

Sayan CHAKRABARTY, *University of Michigan*
Sayan CHAKRABARTY, *University of Michigan*
Elizaveta LEVINA, *University of Michigan*

11:45 Consistency of empirical distributions of sequences of graph statistics in networks with dependent edges [Abstract 228]

Jonathan STEWART, *Florida State University*

03.M2.I55 Modern Methods in High-dimensional Statistics

Venue: Prairie B

Chair : Yisha YAO, Columbia University
Organizer : Cynthia RUSH, Columbia University

10:45 Sparse Autoencoders Demystified: Provable Feature Learning via Adaptive Bias Scheduling [Abstract 246]

Tianhao WANG, *Toyota Technological Institute at Chicago*

11:15 Differentially private penalized M-estimation and inference [Abstract 8]

Marco AVELLA MEDINA, *Columbia University*
Po-Ling LOH, *University of Cambridge*
Zheng LIU,

11:45 Estimating sparse direct effects in multivariate regression with the spike-and-slab LASSO [Abstract 63]

Sameer DESHPANDE, *sameer.deshpande@wisc.edu*
Shen YUNYI, *MIT*
Claudia SOLÍS-LEMUS, *University of Wisconsin–Madison*

03.M2.I56 IMS New Researcher's Group Invited Session

Venue: Prairie C

Chair and Organizer : Satarupa BHATTACHARJEE, University of Florida

10:45 Quantifying common and distinct information in multiomic single-cell data [Abstract 136]

Kevin LIN, *University of Washington*
Kevin LIN, *University of Washington*
Nancy ZHANG, *University of Pennsylvania*
Haoye YANG, *University of Chicago*

11:15 Inference and Learning for Signed Networks Guided by Social Theory [Abstract 231]

Weijing TANG, *Carnegie Mellon University*

11:45 On the Statistical Properties of Generative Adversarial Models for Low Intrinsic Data Dimension [Abstract 34]

Saptarshi CHAKRABORTY, *University of Michigan*
Peter BARTLETT,

Time : **12:15 - 13:30 Lunch** *Venue:* East Campus Dining Center

Plenary Lecture 3 Ryan Tibshirani

Venue: Plains A

Chair : Ashwin PANANJADY, Georgia Tech

13:30 Gradient Equilibrium in Online Learning [Abstract 234]

Ryan TIBSHIRANI, *University of California, Berkeley*
Anastasios ANGELOPOULOS, *University of California, Berkeley*
Michael JORDAN, *University of California, Berkeley*

14:30 - 14:45 Short Break No Coffee

Special Invited Session 5 Jan Hannig, Galin Jones

Venue: Plains A

Chair : Sanjay CHAUDHURI, University of Nebraska-Lincoln

14:45 Fiducial Generative Models [Abstract 93]

Jan HANNIG, *University of North Carolina at Chapel Hill*
Zijie TIAN, *UC Davis*
Thomas C.M. LEE, *tcmlee@ucdavis.edu*

15:30 A Bayesian Generalized Bridge Regression Approach to Covariance Estimation in the Presence of Covariates [Abstract 110]

Galin JONES, *University of Minnesota*
Christina ZHAO, *AbbVie*
Adam ROTHMAN, *University of Minnesota*

03.A1.I57 Enhancing Efficiency and Precision in Clinical Trials: Novel Methods for Outcome Assessment and Data Integration

Venue: Plains B

Chair : Boyu REN, McLean Hospital
Organizer : Himel MALLICK, Cornell University

14:45 Advancing Clinical Dementia Rating (CDR) Analysis: Efficiency Gains Through Item Response Theory [Abstract 135]

Yan LI, *Washington University in St. Louis*

15:15 Improving the Efficiency of Clinical Trials by Making Efficacy Inferences Using Multivariate Endpoints Across Multiple Visits—Lesson Learned from DIAN-TU Platform Trial [Abstract 241]

Guoqiao WANG, *Washington University in St Louis*

15:45 Leveraging External Data for Testing Heterogenous Treatment Effects in Randomized Clinical Trials [Abstract 196]

Boyu REN, *McLean Hospital*
Sandra FORTINI, *Department of Decision Sciences, Bocconi University*
Ventz STEFFEN, *Division of Biostatistics, University of Minnesota*
Lorenzo TRIPPA, *Department of Biostatistics, Harvard T.H. Chan School of Public Health*

03.A1.I58 Innovations in Spatial Statistics

Venue: Plains C

Chair and Organizer : Debjoy THAKUR, Washington University in St. Louis

14:45 Informed MCMC for spatial GLMMs [Abstract 200]

Vivekananda ROY, *Iowa State University*

15:15 Probabilistic Classification and Uncertainty Quantification of Sahara Desert Climate Using Feedforward Neural Networks [Abstract 207]

Indranil SAHOO, *Virginia Commonwealth University*
Stephen TIVENAN, *Virginia Commonwealth University*
Yanjun QIAN, *Virginia Commonwealth University*

03.A1.I59 Frontiers in Learning and Inference in Statistics and AI

Venue: Prairie A

Chair and Organizer : Paromita DUBEY, University of Southern California

14:45 Foundation of Mixture of Experts in Large-Scale Machine Learning Models [Abstract 98]

Nhat HO, *The University of Texas, Austin*

15:15 Sequential Hypothesis Testing via No-Regret Learning [Abstract 242]

Jun-Kun WANG, *UCSD*
Can CHEN, *UCSD*

15:45 Bayesian Joint Additive Factor Models for Multiview Learning [Abstract 151]

Himel MALLICK, *Cornell University*
David DUNSON, *Duke University*
Niccolo ANCESCHI, *Duke University*
Federico FERRARI, *Merck Research Laboratories*

03.A1.I60 Advanced Time Series Analysis Methods and Applications

Venue: Prairie B

Chair and Organizer : Dixon VIMALAJEEWA, University of Nebraska Lincoln

14:45 Goodness of Fit Testing with Saddlepoint Approximation for Degradation Data [Abstract 176]

Lochana PALAYANGODA, *Assistant Professor at University of Nebraska at Omaha*
Hon Keung Tony NG, *Professor at Bentley University*
Aziz GAFUROV, *MS Student at the University of Nebraska at Omaha*

15:15 Enhancing High-Dimensional Time Series Analysis with Envelope Methods [Abstract 97]

Wiranthe HERATH, *Drake University*
Yaser SAMADI, *Southern Illinois University Carbondale*

03.A1.I61 Advances in high-dimensional statistics

Venue: Prairie C

Chair and Organizer : Marco AVELLA MEDINA, Columbia University

14:45 Statistical-computational Trade-offs for Recursive Adaptive Partitioning Estimators [Abstract 123]

Jason KLUSOWSKI, *Princeton University*
Yan Shuo TAN, *National University of Singapore*
Krishna BALASUBRAMANIAN, *University of California, Davis*

15:15 Two-Level SLOPE: Balancing Simplicity and Effectiveness in Adaptive Regularization [Abstract 203]

Cynthia RUSH, *Columbia University*
Zhiqi BU, *Amazon*
Ruijia WU, *SJTU*
Jason KLUSOWSKI, *Princeton University*

15:45 A statistical theory of overfitting for imbalanced classification [Abstract 263]

Kangjie ZHOU, *Columbia University*
Jingyang LYU, *University of Wisconsin, Madison*
Yiqiao ZHONG, *University of Wisconsin, Madison*

03.A1.I62 Innovations and Challenges in Medical Statistics

Venue: Arbor

Chair and Organizer : Manasi SHETH, University of Wisconsin - Whitewater

14:45 Harnessing The Collective Wisdom: Fusion Learning Using Decision Sequences From Diverse Sources [Abstract 11]

Trambak BANERJEE, *University of Kansas*

15:15 Copula-Based Bayesian Model for Detecting Differential Gene Expression [Abstract 27]

Rao CHAGANTY, *Old Dominion University*
Prasansha LIYANAARACHCHI, *University of Sri Jayewardenepura, Sri Lanka*

15:45 Compartmentalization of Discrete Repeated Measures in Patient Reported Outcomes (PROs) [Abstract 219]

Manasi SHETH, *University of Wisconsin - Whitewater*

03.A1.C2 Applications

Venue: Garden

Chair : Maksuda Aktar TOMA, University of Nebraska Lincoln

14:45 Some Mixture models for joint analysis of wind speed and wind direction [Abstract 105]

DEBARGHYA JANA, *Iowa State University*
Arnab HAZRA, *Assistant Professor of Statistics at the Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, Kanpur, India.*

15:00 "Studying algorithmic errors in diagnostic and predictive models in AI-Driven Healthcare Systems: A focus on error detection, mitigation and synthetic data generation" [Abstract 77]

Isaac GBENE, *South Dakota State University*

15:15 Estimation of Parameters of the Truncated Normal Distribution with Unknown Bounds for Robust Methods in Pattern Recognition [Abstract 24]

Dylan BORCHERT, *South Dakota State University*
Semhar MICHAEL, *South Dakota State University*
Christopher SAUNDERS, *South Dakota State University*

15:30 An Empirical Evaluation of GTI Rankings Through Event Data and Media Framing [Abstract 36]

Sagnik CHAKRAVARTY, *University of Maryland College Park*

15:45 100 Years of Pies vs Bars [Abstract 236]

Maksuda Aktar TOMA, *University of Nebraska Lincoln*
Susan VANDERPLAS, *University of Nebraska Lincoln*

16:00 Centile Curve Modelling for Football Performance in Athletic and General Populations [Abstract 61]

Praveen D CHOUGALE, *Indian Institute of Technology Bombay*
Praveen D CHOUGALE, *Indian Institute of Technology Bombay*
Prof. Usha ANANTHAKUMAR, *Shailesh J. Mehta School of Management, I.I.T.Bombay, Powai, Mumbai*

Time : **16:15 - 16:30 Coffee Break** *Venue:* Foyer

03.E1.I63 Recent advances Statistical Genetics

Venue: Plains A

Chair and Organizer : Indranil MUKHOPADHYAY, University of Nebraska Lincoln, USA

16:30 Advances in Heritability Estimation and Reproducible Genomic Inference in Diverse Populations [Abstract 13]

Saonli BASU, *University of Minnesota*

17:00 Stability and Shrinkage Selection in High Dimensional Logistic Regression [Abstract 52]

Jennifer CLARKE, *University of Nebraska-Lincoln*
Bertrand CLARKE, *University of Nebraska-Lincoln*
Caleb HOLMBECK, *University of Nebraska-Lincoln*
Laura KRESTY, *University of Michigan*

17:30 Prediction of Feed Efficiency Traits in Beef Cattle Using Host Genomic and Metagenomic Sequence Data [Abstract 225]

Matt SPANGLER, *University of Nebraska-Lincoln*
Matthew SPANGLER, *University of Nebraska-Lincoln*
Andrew LAKAMP, *University of Nebraska-Lincoln*
Samodha FERNANDO, *University of Nebraska-Lincoln*

Stat Bowl Stat Bowl

Venue: Plains B

Organizers : *Rajarshi DE, Emporia State University* and *Ananda SEN, University of Michigan Ann Arbor*

03.E1.I64 Statistical and Computational Advances in Complex Decision-Making

Venue: Plains C

Chair : Runmin WANG, Texas A&M University
Organizer : Hao CHEN, University of California, Davis

16:30 Optimal Sequential Recommendation Systems [Abstract 116]

Mina KARZAND, *UC Davis*
Mina KARZAND, *Department of Statistics, UC Davis*
Guy BRESLER, *EECS Department, MIT*

17:00 A Graph-based Approach to Estimating the Number of Clusters in High-dimensional Settings [Abstract 51]

Lynna CHU, *Iowa State University*
Yichuan BAI, *Iowa State University*
Lynna CHU, *Iowa State University*

17:30 Empirical Error Estimates for Graph Sparsification [Abstract 138]

Miles LOPES, *UC Davis*
Siyao WANG,

03.E1.I65 Advances in approximate Bayesian learning

Venue: Prairie A

Chair : Shuang ZHOU, Arizona State University
Organizer : Anirban BHATTACHARYA, Texas A&M University

16:30 Statistical Guarantees for Semi-Implicit Variational Inference [Abstract 187]

Sean PLUMMER, *University of Arkansas*

17:00 Approximating Distributions via Deep Generative Models: Theory, Limitations and Directions [Abstract 230]

Edric TAM, *Stanford University*
Edric TAM, *Stanford University*
David DUNSON, *Duke University*

17:30 Addressing Antidiscrimination with Variational Inference [Abstract 264]

Shuang ZHOU, *Arizona State University*
Lydia GABRIC, *Arizona State University*
Kenneth ZHOU, *University of Waterloo*

03.E1.I66 Advances in Statistical Methods for Training, Serving, and Evaluating Large Language Models

Venue: Prairie B

Chair and Organizer : Rohit PATRA, Linkedin Inc

16:30 Training of Efficient LLMs for Industry Applications: Methods and Insights [Abstract 14]

Kayhan BEHDIN, *LinkedIn*

17:00 **TBD** [Abstract 185]

Rohit PATRA, *Linkedin Inc*

03.E1.I67 Advancements in Statistical Modeling for Complex Data: Microbiome Associations, Low-Rank Matrix Models, and Network Functional Connectivity

Venue: Prairie C

Chair and Organizer : Sharmistha GUHA, Texas A&M University

16:30 **Coherence-free Entrywise Estimation of Eigenvectors in Low-rank Signal-plus-noise Matrix Models** [Abstract 129]

Keith LEVIN, *University of Wisconsin, Madison*

17:00 **Structured Bayesian Variable Selection for Microbiome Compositional Data Using Graph-Guided Shrinkage** [Abstract 206]

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

17:30 **Bayes in Multi-Layer Networks** [Abstract 88]

Sharmistha GUHA, *Texas A&M University*

03.E1.I68 Statistical and Computational Methods for Complex Data

Venue: Arbor

Chair : Changbo ZHU, University of Notre Dame

Organizer : Satarupa BHATTACHARJEE, University of Florida

16:30 **Belted and Ensembled Neural Network for Linear and Nonlinear Sufficient Dimension Reduction** [Abstract 232]

Yin TANG, *Pennsylvania State University*
Bing LI, *Pennsylvania State University*

17:00 **Elastic Net-Based Variable Selection for Fréchet Regression in RKHS** [Abstract 252]

Haoyi YANG, *Department of Statistics, The Pennsylvania State University, USA*
Bing LI,
Lingzhou XUE,
Satarupa BHATTACHARJEE,

17:30 **Doubly robust estimation of causal effects for random object outcomes with continuous treatments** [Abstract 18]

Satarupa BHATTACHARJEE, *University of Florida*
Bing LI, *Pennsylvania State University*
Lingzhou XUE, *Pennsylvania State University*
Xiao WU, *Columbia University*

Sunday June 15

04.M1.I69 Recent developments in Statistical Applications

Venue: Plains A

Chair : Reka HOWARD, University of Nebraska-Lincoln
Organizer : Abishek BHATTACHARJEE, Pfizer

9:00 **Bayesian Analysis of Space Sustainability issues** [Abstract 167]

Ujjal MUKHERJEE, *University of Illinois Urbana Champaign*
Ujjal MUKHERJEE, *University of Illinois Urbana-Champaign*
Snigdhansu CHATTERJEE, *University of Maryland Baltimore County*

9:30 **Half-orthogonalized Neural Network for Estimating Yield Response Function Using On-farm Experiment Data** [Abstract 157]

Taro MIENO, *University of Nebraska Lincoln*
Mona MOISAVI, *University of Nebraska Lincoln*
David BULLOCK, *University of Illinois at Urbana Champaign*

10:00 **Allowing Negative Variance Component Estimates in REML: Inferential Consequences for Fixed Effects and Type I Error Control** [Abstract 100]

Reka HOWARD, *University of Nebraska-Lincoln*
Bipin POUDEL, *University of Nebraska-Lincoln*
Nora BELLO, *USDA*
Walt STROUP, *University of Nebraska-Lincoln*

04.M1.I70 Causal inference in complex settings

Venue: Plains B

Chair : Rajarshi MUKHERJEE, Harvard University
Organizer : Peng DING, University of California Berkeley

9:00 **Design-based inference for incomplete block designs** [Abstract 182]

Nicole PASHLEY, *Rutgers University*
Taehyeon KOO,

9:30 **High-dimensional moderated mediation analysis with heredity** [Abstract 265]

Wen ZHOU, *New York University*
Zhang ZIFENG, *Colorado State University*
Fan YANG, *Tsinghua University*
Peng DING, *UC Berkeley*

10:00 **Method-of-Moments Inference for GLMs and Doubly Robust Functionals under Proportional Asymptotics** [Abstract 164]

Rajarshi MUKHERJEE, *Harvard University*
Xingyu CHEN, *Shanghai Jiao Tong University*
Liu LIN , *Shanghai Jiao Tong University*

04.M1.I71 Modern statistical methods, with applications in the biomedical sciences

Venue: Plains C

Chair and Organizer : Cynthia RUSH, Columbia University

9:00 **Microbiome Data Integration via Shared Dictionary Learning** [Abstract 245]

Shulei WANG, *University of Illinois Urbana-Champaign*
Bo YUAN, *University of Illinois Urbana-Champaign*

9:30 **Improved automated cryo-EM structure determination via diffusion model** [Abstract 253]

Yisha YAO, *Columbia University*
Xiaoyu FANG, *Columbia University*
Sheng CHEN, *Sun Yat-sen University*

10:00 **Learning Joint and Individual Structure in Network Data with Covariates** [Abstract 6]

Jesús ARROYO, *Texas A&M University*
Carson JAMES, *Texas A&M University*
Dongbang YUAN, *Meta*
Irina GAYNANOVA, *University of Michigan*

Time : **10:30 - 10:45 Coffee Break** *Venue:* **Foyer**

04.M2.I72 Recent advances in high-dimensional modeling

Venue: Plains A

Chair : Partha SARKAR, Florida State University
Organizer : Sayantan BANERJEE, Indian Institute of Management Indore

10:45 **Mixed Model Trace Regression** [Abstract 227]

Sanvesh SRIVASTAVA, *The University of Iowa*
Ian HULTMAN, *The University of Iowa*

11:15 **Bayesian Semiparametric Functional Mixed Effects Drift-Diffusion Models for Cognitive Control Leveraging Stop-signal Tasks** [Abstract 38]

Noirrit Kiran CHANDRA, *The University of Texas at Dallas*
Farabi Raihan SHUVO, *The University of Texas at Dallas*
Stacie WARREN, *The University of Texas at Dallas*

11:45 **Misspecified Yet Credible: A Generalized Bayes Framework for Uncertainty Quantification in High-Dimensional Bayesian Vector Autoregressive Models** [Abstract 210]

Partha SARKAR, *Florida State University*
Ray BAI, *University of South Carolina*

04.M2.I73 Bayesian Inference and Uncertainty Quantification in Regression, Dynamic Systems, and Inverse Problems

Venue: Plains B

Chair and Organizer : Paromita BANERJEE, John Carroll University

10:45 **Bayesian Ordinal Network Meta-Regression under General Links with Applications to Crohn's Disease** [Abstract 91]

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

11:15 **Bayesian inference for COVID-19 transmission dynamics using a modified SEIR model** [Abstract 159]

Anirban MONDAL, *Case Western Reserve University*
Kai YIN, *Case Western Reserve University*
Paromita BANERJEE, *John Carroll University*
David GURARIE, *Case Western Reserve University*

11:45 **Deep Learning based surrogate models in Statistical Inverse Problems** [**Abstract 1**]

Anuj ABHISHEK, *Case Western Reserve University*
Sudeb MAJEE, *UNCC*
Thilo STRAUSS, *XJTLU, China*
Taufiquar KHAN, *UNCC*

04.M2.I74 Bayesian and Empirical Methods for Prediction, Inference, and Signal Detection

Venue: Plains C
Chair : Saptarshi CHAKRABORTY, University at Buffalo
Organizer : Abishek BHATTACHARJEE, Pfizer

10:45 **Beyond the Odds: Fitting Logistic Regression with Missing Data in Small Samples** [**Abstract 188**]

Vivek PRADHAN, *Pfizer Inc., Cambridge, MA 02139, USA*

11:15 **Statistical modeling and prediction of patient recruitment in multicenter clinical trials** [**Abstract 208**]

Srijata SAMANTA, *Bristol Myers Squibb*

11:45 **Flexible Empirical Bayesian Approaches to Pharmacovigilance for Simultaneous Signal Detection and Signal Strength Estimation in Spontaneous Reporting Systems Data** [**Abstract 35**]

Saptarshi CHAKRABORTY, *University at Buffalo*
Yihao TAN, *University at Buffalo*
Marianthi MARKATOU, *University at Buffalo*

Time : **12:15 - 13:30** **Lunch** *Venue:* East Campus Dining Center

Abstracts

1. Deep Learning based surrogate models in Statistical Inverse Problems

[04.M2.I73, (page 38)]

Anuj ABHISHEK, *Case Western Reserve University*
Sudeb MAJEE, *UNCC*
Thilo STRAUSS, *XJTLU, China*
Taufiquar KHAN, *UNCC*

Neural operators such as Deep Operator Networks (DeepONet) and Convolutional Neural Operators (CNO) have been shown to be fairly useful in approximating an operator between two function spaces. In this talk, we at first show that they can be used to approximate operators that are maps between more general Banach spaces (not necessarily just function spaces) and which appear in various important medical imaging problems. Following recent developments in the field, we derive universal approximation theorem type results for two different network implementations that are used for learning the types of operators that turn up in imaging modalities such as EIT, DOT and QPAT. We then show how these operator learning frameworks may be used for direct inversion as well as may be used as surrogate models for the likelihood evaluation in Bayesian inversion. This is based on joint works with Thilo Strauss (Xi'an Jiaotong-Liverpool University) and Taufiquar Khan and Sudeb Majee (UNC Charlotte).

2. Network models for spatial transcriptomics data

[03.M1.I52, (page 28)]

Satwik ACHARYYA, *University of Alabama at Birmingham*

Network models are powerful tools to investigate complex dependence structures in high throughput genomic datasets. They allow for holistic, systems-level view of the various biological processes, for intuitive understanding and coherent interpretations. However, most existing network or graphical models are developed under assumptions of homogeneity of samples and are not readily amenable to modeling spatial heterogeneity which often manifests in spatial genomics data. In this talk, I will discuss two spatial network models focusing on spatially varying covariance and precision matrices. (I) SpaceX (spatially dependent gene co-expression network) is a Bayesian methodology to identify both shared and cluster-specific co-expression networks across genes. (II) Spatial Graphical Regression (SGR) is a flexible approach based on graphical regression that enables spatially varying graphs over the spatial domain of the tissue. The framework incorporates multiple spatial covariates and provides a non-linear functional mapping between the spatial domain and the precision matrices. All the approaches are illustrated by using case studies from cancer genomics.

3. Estimation and Inference in Tensor Mixed-Membership Blockmodels

[02.M1.I21, (page 15)]

Joshua AGTERBERG, *University of Illinois Urbana-Champaign*
Anru ZHANG, *Duke University*

Higher-order multiway data is ubiquitous in machine learning and statistics and often exhibits community-like structures, where each component (node) along each different mode has a community membership associated with it. In this talk we propose the tensor mixed-membership blockmodel, a generalization of the tensor blockmodel positing that memberships need not be discrete, but instead are convex combinations of latent communities. We first study the problem of estimating community memberships, and we show that a tensor generalization of a matrix algorithm can consistently estimate communities at a rate that improves relative to the matrix setting, provided one takes the tensor structure into account. Next, we study the problem of testing whether two nodes have the same community memberships, and we show that a tensor analogue of a matrix test statistic can yield consistent testing with a tighter local power guarantee relative to the matrix setting. If time permits we will also examine the performance of our estimation procedure on flight data.

4. Aligning Multiple Inhomogeneous Random Graphs: Fundamental Limits of Exact Recovery

[Student Paper Competition 2, (page 5)]

Taha AMEEN, *University of Illinois Urbana-Champaign*
Bruce HAJEK, *University of Illinois Urbana-Champaign*

This work studies fundamental limits for recovering the underlying correspondence among multiple correlated graphs. In the setting of inhomogeneous random graphs, we present and analyze a matching algorithm: first partially match the graphs pairwise and then combine the partial matchings by transitivity. Our analysis yields a sufficient condition on the problem parameters to exactly match all nodes across all the graphs. In the setting of homogeneous (Erdos-Renyi) graphs, we show that this condition is also necessary, i.e. the algorithm works down to the information theoretic threshold. This reveals a sce-

nario where exact matching between two graphs alone is impossible, but leveraging more than two graphs allows exact matching among all the graphs. Converse results are also given in the inhomogeneous setting and transitivity again plays a role. Along the way, we derive independent results about the k-core of inhomogeneous random graphs.

5. Modeling Spatio-Temporal Environmental Processes and Emerging Crises under Data Sparsity

[02.A1.I33, (page 21)]

Ali ARAB, *Georgetown University, Department of Mathematics and Statistics*

Modeling the dynamics of spatio-temporal processes is often challenging and this is exacerbated under data sparsity (often the case in early stages of a process). For example, modeling the dynamics of a vector-borne infectious disease at early stages is very challenging due to data sparsity (as well as potential lack of knowledge regarding the disease dynamics itself); this is an important issue for modeling emerging and re-emerging epidemics, or emerging climate crises. Moreover, data sparsity may also result in inefficient inference and ineffective prediction for such processes. This is a common issue in modeling rare or emerging ecological, environmental, epidemiological, and social processes that are new or uncommon in specific areas, specific time periods, or those conditions that are hard to detect. Consequently, due to the urgency of modeling these processes in many situations (e.g., in a crisis situation), often one limited predictor data to use either because of lack of knowledge about the process or the need for fine resolution predictor data. For example, modeling the dynamics of climate-driven human migration may be quite complex to model (in particular, when a crisis occurs and there are abrupt migration in/out flows). Classic models that are commonly used in these areas often fall short of modeling such events and are unable to provide reliable inference and reasonable or accurate forecasts. Also, the factors that are linked with migration processes are often related to long term migration and/or only available at spatially and temporally aggregated level. Here, we discuss strategies for dealing with some of the statistical issues of modeling dynamics under data sparsity including: utilizing blended data (i.e., both conventional and organic data sources), considering a mechanistic science-based modeling framework to model the dynamics of a spatio-temporal based on zero-modified hierarchical modeling approaches, and implementing improved parameter estimation and forecasting through transfer learning. Case studies will be discussed.

6. Learning Joint and Individual Structure in Network Data with Covariates

[04.M1.I71, (page 37)]

Jesús ARROYO, *Texas A&M University*
Carson JAMES, *Texas A&M University*
Dongbang YUAN, *Meta*
Irina GAYNANOVA, *University of Michigan*

Datasets consisting of a network and covariates associated with its vertices have become ubiquitous. One problem pertaining to this type of data is to identify information unique to the network, information unique to the vertex covariates and information that is shared between the network and the vertex covariates. Existing techniques for network data and vertex covariates focus on capturing structure that is shared but are usually not able to differentiate structure that is unique to each dataset. This work formulates a low-rank model that simultaneously captures joint and individual information in network data with vertex covariates. A two-step estimation procedure is proposed, composed of an efficient spectral method followed by a refinement optimization step. Theoretically, we show that the spectral method is able to consistently recover the joint and individual components under a general signal-plus-noise model. Simulations and real data examples demonstrate the ability of the methods to recover accurate and interpretable components. Application of the methodology to a food trade network and brain network data are discussed.

7. Bayesian Hierarchical Modeling for Extremes

[01.A1.I11, (page 11)]

Sakshi ARYA, *Case Western Reserve University*
Snigdhanshu CHATTERJEE, *UMBC*
Vishal SUBEDI, *UMBC*
Dietz LINDSEY, *University of Minnesota*

Predicting hurricane-related monetary losses is vital for disaster planning, relying on storm attributes like wind speed and pressure. Despite existing efforts, there's a need for reliable tools to estimate losses. This work aims to jointly model past hurricane damages and storm characteristics using specialized statistical methods. We propose a novel approach using Generalized Extreme Value distributions in a Bayesian framework, extending it to seasonal predictions. We incorporate major climate indices as covariates and conduct robustness checks. We present some theoretical results for our proposed methodol-

ogy and show that our method successfully predicts damages for Atlantic cyclones during 2016-22, aligning closely with actual costs. Seasonal analysis estimates average annual damages for the United States and demonstrates high accuracy in predicting storm damages.

8. Differentially private penalized M-estimation and inference

[03.M2.I55, (page 29)]

Marco AVELLA MEDINA, *Columbia University*
Po-Ling LOH, *University of Cambridge*
Zheng LIU,

We propose a noisy composite gradient descent algorithm for differentially private statistical estimation in high dimensions. We begin by providing general rates of convergence for the parameter error of successive iterates under assumptions of local restricted strong convexity and local restricted smoothness. Our analysis is local, in that it ensures a linear rate of convergence when the initial iterate lies within a constant-radius region of the true parameter. At each iterate, multivariate Gaussian noise is added to the gradient in order to guarantee that the output satisfies Gaussian differential privacy. We then derive consequences of our theory for linear regression and mean estimation. Motivated by M-estimators used in robust statistics, we study loss functions which downweight the contribution of individual data points in such a way that the sensitivity of function gradients is guaranteed to be bounded, even without the usual assumption that our data lie in a bounded domain. We prove that the objective functions thus obtained indeed satisfy the restricted convexity and restricted smoothness conditions required for our general theory. We then show how the private estimators obtained by noisy composite gradient descent may be used to obtain differentially private confidence intervals for regression coefficients, by leveraging work in Lasso debiasing proposed in high-dimensional statistics.

9. Modeling Spatial Extremes using Non-Gaussian Spatial Autoregressive Models via Convolutional Neural Networks

[03.M1.I46, (page 26)]

Soutir BANDYOPADHYAY, *Colorado School of Mines*
Sweta RAI, *Colorado School of Mines*
Douglas NYCHKA, *Colorado School of Mines*

Data derived from remote sensing or numerical simulations often have a regular gridded structure and are large in volume making it challenging to find accurate spatial models that can fill in missing grid cells or simulate the process effectively, especially when there is spatial heterogeneity and heavy-tailed marginal distributions. To overcome this issue, in this work we present a spatial autoregressive (SAR) modeling framework, which maps observations at a location and its neighbors to spatially independent random variables. This is a flexible modeling approach and well-suited for non-Gaussian fields, providing simpler interpretability. In particular, in this work, we consider the SAR model with Generalized Extreme Value (GEV) distribution innovations to combine the observation at a central grid location with its neighbors, capturing extreme spatial behavior based on the heavy-tailed innovations. While these models are fast to simulate exploiting the sparsity of the key matrices in the computations, the maximum likelihood estimation of the parameters is slow because the likelihood is intractable and hence difficult to optimize. To overcome this, we train a convolutional neural network (CNN) on a large training set that spans a useful parameter space, then using the trained network for fast estimation. We apply this model to analyze annual maximum precipitation derived from ERA-Interim-driven WRF simulations obtained from the NA-CORDEX project, allowing us to explore spatial extremal behavior across North America.

10. Robust K-means Clustering based on Density Power Divergence Measure

[02.M1.I23, (page 16)]

Paromita BANERJEE, *John Carroll University*
Anirban MONDAL, *Case Western Reserve University*
Abhijit MANDAL, *University of Texas at El Paso*

We introduce a robust clustering method as a modification of the widely used partition-optimization algorithm, K-means. The traditional K-means algorithm partitions observations into K clusters by minimizing within-cluster variances. However, it is susceptible to the influence of outliers, which can distort the estimation of cluster means and distance metrics, resulting in inappropriate clustering allocations. In our proposed robust K-means clustering method, we estimate cluster centers and covariance matrices using density power divergence (DPD) measures. The tuning parameter in the DPD measure allows adjustment for outlier contamination, providing optimal robust estimates for cluster centers, covariance matrices, and clusters. To address the limitations of the Euclidean distance measure, particularly when clusters have heterogeneous and el-

liptical shapes, we employ Mahalanobis distance for computing the distance between each point and cluster center. The efficacy of our proposed method is demonstrated through its application to simulated data, showcasing its superiority over existing methods based on established evaluation metrics. Furthermore, we apply the methodology to two real datasets: 1) clustering Irish data to identify similar species and 2) clustering countries based on Covid-19 case fatality rate and infection rate, while examining the impact of socio-economic and environmental factors within those clusters.

11 . Harnessing The Collective Wisdom: Fusion Learning Using Decision Sequences From Diverse Sources

[03.A1.I62, (page 32)]

Trambak BANERJEE, *University of Kansas*

Learning from the collective wisdom of crowds is related to the statistical notion of fusion learning from multiple data sources or studies. However, fusing inferences from diverse sources is challenging since cross-source heterogeneity and potential data-sharing complicate statistical inference. Moreover, studies may rely on disparate designs, employ myriad modeling techniques, and prevailing data privacy norms may forbid sharing even summary statistics across the studies for an overall analysis. We propose an Integrative Ranking and Thresholding (IRT) framework for fusion learning in multiple testing. IRT operates under the setting where from each study a triplet is available: the vector of binary accept-reject decisions on the tested hypotheses, its False Discovery Rate (FDR) level and the hypotheses tested by it. Under this setting, IRT constructs an aggregated and nonparametric measure of evidence against each null hypotheses, which facilitates ranking the hypotheses in the order of their likelihood of being rejected. We show that IRT guarantees an overall FDR control if the studies control their respective FDR at the desired levels. IRT is extremely flexible, and a comprehensive numerical study demonstrates its practical relevance for pooling inferences. A real data illustration and extensions to alternative forms of Type I error control are discussed.

12. TBD

[01.A1.I10, (page 11)]

Sanjib BASU, *University of Illinois at Chicago*

TBA

13. Advances in Heritability Estimation and Reproducible Genomic Inference in Diverse Populations

[03.E1.I63, (page 33)]

Saonli BASU, *University of Minnesota*

Understanding the genetic architecture of complex traits and diseases depends on accurate heritability estimation, the quantification of phenotypic variance explained by genetic factors. While methods such as GREML and LD Score Regression have been instrumental, challenges persist in accounting for population structure, rare variants, and high-dimensional data. This talk will highlight recent methodological advances in heritability estimation, with a particular emphasis on scalable approaches designed for large, ancestrally diverse populations.

In parallel, I will address the growing need for localized, reproducible genomic data infrastructures that support transparent and efficient analysis. At the University of Minnesota, the Genomic Data Commons (UMN-GDC) provides a standards-driven, cloud-compatible environment for secure, in-place analysis of genomic data. I will describe how this platform promotes scientific rigor through data harmonization, metadata standardization, embedded version control, and containerized workflows. Special focus will be given to modules designed for benchmarking reproducibility and for deploying heritability estimation pipelines across heterogeneous datasets. Together, these innovations advance a reproducible, scalable, and equitable ecosystem for genomic discovery.

14. Training of Efficient LLMs for Industry Applications: Methods and Insights

[03.E1.I66, (page 34)]

Kayhan BEHDIN, *LinkedIn*

Large language models (LLMs) have demonstrated remarkable performance across a wide range of industrial applications, from search and recommendations to generative tasks. Although scaling laws indicate that larger models generally yield better generalization and performance, their substantial computational requirements often render them impractical for many real-world scenarios at scale. In this work, we present methods and insights for compressing LLMs into Small Language Models (SLMs) that maintain the generalization performance of the LLM, while requiring less inference compute resources. We focus on two key techniques: (1) knowledge distillation and (2) model compression via quantization and pruning. We discuss the methodological aspects of these techniques. Additionally, we share the lessons we have learned from applying model compression

techniques in a large-scale professional social network.

15. Frequency Domain Resampling for Gridded Spatial Data

[02.M1.C1, (page 16)]

Souvick BERA, *Colorado School of Mines*
Daniel J. NORDMAN, *Iowa State University*
Soutir BANDYOPADHYAY, *Colorado School of Mines*

In frequency domain analysis for spatial data, spectral averages based on the periodogram often play an important role in understanding spatial covariance structure, but also have complicated sampling distributions due to complex variances from aggregated periodograms. In order to nonparametrically approximate these sampling distributions for purposes of inference, resampling can be useful, but previous developments in spatial bootstrap have faced challenges in the scope of their validity, specifically due to issues in capturing the complex variances of spatial spectral averages. As a consequence, existing frequency domain bootstraps for spatial data are highly restricted in application to only special processes (e.g. Gaussian) or certain spatial statistics. To address this limitation and to approximate a wide range of spatial spectral averages, we propose a practical hybrid-resampling approach that combines two different resampling techniques in the forms of spatial subsampling and spatial bootstrap. Subsampling helps to capture the variance of spectral averages while bootstrap captures the distributional shape. The hybrid resampling procedure can then accurately quantify uncertainty in spectral inference under mild spatial assumptions. Moreover, compared to the more studied time series setting, this work fills a gap in the theory of subsampling/bootstrap for spatial data regarding spectral average statistics.

16. Non-parametric mixture models for covariance function estimation

[02.E1.I43, (page 24)]

Stephen BERG, *Penn State Statistics*
Hyebin SONG,

I will introduce nonparametric, shape-constrained estimation for covariance functions, with an emphasis on a shape-constrained weighted least squares estimator of the autocovariance sequence from a reversible Markov chain. The estimator will be shown to lead to strongly consistent estimates of the asymptotic variance of the sample mean from an MCMC sample, as well as to L_2 consistent estimates of the autocovariance sequence. An algorithm for computing our estimator will be presented, and some empirical applications will be shown. The proposed shape-constrained estimator exploits a mixture representation of the autocovariance sequence from a reversible Markov chain. Similar mixture representations exist for stationary covariance functions in spatial statistics, including for the Matérn covariance as a special case, and I will highlight some extensions of shape-constrained approaches for estimating covariance functions in spatial statistics.

17. "Riten Da : sometimes, it's ok!"

[02.M1.I20, (page 15)]

Sagnik BHADURY, *University of Michigan*

18. Doubly robust estimation of causal effects for random object outcomes with continuous treatments

[03.E1.I68, (page 35)]

Satarupa BHATTACHARJEE, *University of Florida*
Bing LI, *Pennsylvania State University*
Lingzhou XUE, *Pennsylvania State University*
Xiao WU, *Columbia University*

Causal inference is central to statistics and scientific discovery, enabling researchers to identify cause-and-effect relationships beyond associations. While traditionally studied within Euclidean spaces, contemporary applications increasingly involve complex, non-Euclidean data structures that reside in abstract metric spaces, known as random objects, such as images, shapes, networks, and distributions. This paper introduces a novel framework for causal inference for continuous treatments in such settings. To address the challenges posed by the lack of linear structures, we leverage Hilbert space embeddings of the metric spaces to facilitate Fréchet mean estimation and causal effect mapping. Motivated by an environmental study on the impact of fine particulate matter ($\geq 2.5\mu m$) air pollution on age-at-death distributions across U.S. counties, we propose a nonparametric, doubly-debiased inference approach for random objects with continuous treatments. Our framework can accommodate moderately high-dimensional vector-valued confounders and derives efficient influence functions for estimation to ensure both robustness and interpretability. We establish rigorous asymptotic properties of the cross-fitted estimators and employ conformal inference techniques for counterfactual outcome prediction. Validated through numerical experiments and applied to real-world environmental data, our framework extends causal inference methodologies to complex data structures, broadening its applicability across scientific disci-

plines.

19 . Inference on Network Structures of Hypergraphs under Edge Exchangeability

[Student Poster Competition, (page 6)]

Ayoushman BHATTACHARYA, *Department of Statistics and Data Science, Washington University in St. Louis*
Nilanjan CHAKRABORTY, *Department of Mathematics and Statistics, Missouri University of Science and Technology*
Robert LUNDE, *Department of Statistics and Data Science, Washington University in St. Louis*

In statistical network analysis, it is common practice to consider models for binary adjacency matrices that satisfy vertex exchangeability. However, these models may fail to capture important properties of the data generating process when the fundamental units are interactions rather than nodes. We consider the problem of statistical inference for subgraph counts under an exchangeable hyperedge model. We introduce various classes of subgraph statistics for hypergraphs, and develop inferential tools for a notion of subgraph frequency that take into account the multiplicity of an edge. We further show that a certain subclass of these subgraph statistics is robust to the deletion of low-degree nodes, facilitating inference in settings where low-degree nodes are more likely to be missing. We also consider a more traditional notion of subgraph frequency that does not take into account multiplicity. We demonstrate that while statistical inference for these statistics is possible in certain cases, in many cases, a proper limiting distribution may not even exist. We study the finite sample properties of our procedures in both simulation studies and real-world datasets. For our data analysis, we collect new hypergraph datasets involving academic and movie collaborations, and we find that our inferential tools for hypergraphs appear to have more power to distinguish between networks than traditional approaches based on binary adjacency matrices.

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

[01.M2.I4, (page 4)]

Sudipta BHATTACHARYA, *Daiichi Sankyo, Inc.*
Jyotirmoy DEY, *Regeneron*

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 is an amalgamation of these two ideas to investigate the contribution of a specific component of a regimen comprising multiple treatment phases. Different variants of the method are provided and indices of contribution of a treatment phase to the overall benefit of a regimen are constructed that facilitates interpretation of results.

21 . Time-stamped Networks and Hawkes Process with Missing Data

[03.M2.I53, (page 29)]

Sharmodeep BHATTACHARYYA, *Oregon State University*
Sarah EMERSON, *Oregon State University*
Robert TRANGUCCI, *Oregon State University*
Jingtian YU, *Oregon State University*

Temporal or time-stamped networks are characterized by the triplets (i,j,t), which capture the presence of an edge between nodes I and j at time point t. A point process formulation of the edge formation events encounters a missing value problem on account of edge sparsity. Specifically, we study the Hawkes process as the point process of edge generation event. The Hawkes process is a self-exciting stochastic process where event intensity depends on past occurrences, making estimations with incomplete observations a challenge as unobserved events still affect the dynamics. Previous studies have investigated estimation problems under certain structures of missing mechanisms. In this work, we describe a general mechanism where the probability of an event being missing does not depend on the event's position in the time sequence. We develop a likelihood-based estimation approach that incorporates proposal steps tailored to accommodate the missing mechanism, ensuring robust handling of incomplete data scenarios.

22 . Inference in Federated Learning: limit theorems and applications

[02.M1.C1, (page 16)]

Soham BONNERJEE, *University of Chicago*
Sayar KARMAKAR, *University of Florida*

We start by analyzing M-estimation in the general DFL algorithm from a time series perspective. This particular pov is helpful in establishing sharp Gaussian approximations, which are applied to perform bootstrap-based inference.

23 . Higher-order estimators of time-varying effects in anisotropic smoothness models

[02.M2.I26, (page 18)]

Matteo BONVINI, *Rutgers University*
Edward KENNEDY, *Carnegie Mellon University*
Luke KEELE, *University of Pennsylvania*

The general theory of higher-order influence functions (HOIF) has been successfully applied to several pathwise differentiable parameters arising in causal inference, such as the expected conditional covariance and the treatment-specific mean. Such theory has been shown to yield minimax optimal estimators in certain nonparametric models, e.g., those indexed by smooth nuisance parameters. More recently, minimax optimal, higher-order estimators have been derived for some non-pathwise differentiable causal parameters, an example of which is the conditional average treatment effect. In this work, we aim to extend the application of HOIF theory to causal parameters defined by a time-varying treatment. As a leading example, we consider the two-time point case g-formula functional in an anisotropic smoothness model where the nuisance functions can depend more smoothly on certain covariates. We also consider even more structured models, such as additive ones. In each setting, we design a higher-order estimator and calculate its bias and variance, and for some of them, we show that the convergence rates established are minimax optimal. We complement our findings with simulations.

24 . Estimation of Parameters of the Truncated Normal Distribution with Unknown Bounds for Robust Methods in Pattern Recognition

[03.A1.C2, (page 33)]

Dylan BORCHERT, *South Dakota State University*
Semhar MICHAEL, *South Dakota State University*
Christopher SAUNDERS, *South Dakota State University*

Estimators of parameters of truncated distributions, namely the truncated normal distribution, have been widely studied for a known truncation region. There is also literature for estimating the unknown bounds for known parent distributions. However, to our knowledge, there are no works that accommodate both parameter and bound estimation of the truncated normal distribution. In this work, we develop a novel algorithm under the expectation-solution (ES) framework, which is an iterative method of solving nonlinear estimating equations, to estimate both the bounds and the location and scale parameters of the parent normal distribution utilizing theory of best linear unbiased estimates from location-scale families of distribution and unbiased minimum variance estimation of truncation regions. The computational aspects of the algorithm and asymptotic properties or the resulting estimators are discussed, and an illustrative example of the utility of the parameter estimates is shown in an application to pattern recognition with real data.

25 . Nonparametric Within-Between Models

[Student Paper Competition 2, (page 5)]

Soumyabrata BOSE, *University of Texas at Austin*
Antonio R. LINERO, *University of Texas at Austin*
Jared S. MURRAY, *University of Texas at Austin*

The within-between (WB) model is a robust approach that addresses the constraints inherent in both fixed effects (FE) and random effects (RE) models by distinctly modeling within-group and between-group effects. This paper introduces a nonparametric extension of the within-between model for the analysis of hierarchical data, with the data subsequently modeled in a Bayesian nonparametric fashion. Our extension permits flexible nonlinear relationships while preserving the interpretability benefits of the linear within-between framework. We establish theoretical guarantees on posterior concentration rates under appropriate conditions and present a framework for deriving interpretable summaries of the nonparametric effects using surrogate models. Through simulation studies, we demonstrate the superior performance of our approach compared to existing methods, including linear fixed effects, random effects, and standard BART extensions (BART-RE), particularly when the true relationships are nonlinear. We illustrate the practical applicability of our method through its application to the National Education Longitudinal Study, wherein we analyze student dropout status while accounting for both student-level and school-level effects.

26. Minimax Optimality of the Probability Flow ODE for Diffusion Models

[02.M2.I29, (page 19)]

Changxiao CAI, *University of Michigan*
Gen LI, *The Chinese University of Hong Kong*

Score-based diffusion models have become a foundational paradigm for modern generative modeling, demonstrating exceptional capability in generating samples from complex high-dimensional distributions. Despite the dominant adoption of probability flow ODE-based samplers in practice due to their superior sampling efficiency and precision, rigorous statistical guarantees for these methods have remained elusive in the literature. This work develops the first end-to-end theoretical framework for deterministic ODE-based samplers that establishes near-minimax optimal guarantees under mild assumptions on target data distributions. Specifically, focusing on subgaussian distributions with β-Hölder smooth densities for $\beta \leq 2$, we propose a smooth regularized score estimator that simultaneously controls both the L^2 score error and the associated mean Jacobian error. Leveraging this estimator within a refined convergence analysis of the ODE-based sampling process, we demonstrate that the resulting sampler achieves the minimax rate in total variation distance, modulo logarithmic factors. Notably, our theory comprehensively accounts for all sources of error in the sampling process and does not require strong structural conditions such as density lower bounds or Lipschitz/smooth scores on target distributions, thereby covering a broad range of practical data distributions.

27. Copula-Based Bayesian Model for Detecting Differential Gene Expression

[03.A1.I62, (page 32)]

Rao CHAGANTY, *Old Dominion University*
Prasansha LIYANAARACHCHI, *University of Sri Jayewardenepura, Sri Lanka*

DNA is a fundamental genetic material in all living organisms, contains thousands of genes, but only a subset exhibits differential expression and play a crucial role in diseases. Microarray technology has revolutionized the study of gene expression, with two primary types available for expression analysis: spotted cDNA arrays and oligonucleotide arrays. This research focuses on the statistical analysis of data from spotted cDNA microarrays. Numerous models have been developed to identify differentially expressed genes based on the red and green fluorescence intensities measured by these arrays. We propose a novel approach using a Gaussian copula model to characterize the joint distribution of red and green intensities, effectively capturing their dependence structure. Given the right-skewed nature of the intensity distributions, we model the marginal distributions using gamma distributions. Differentially expressed genes are identified using the Bayes estimate under our proposed copula framework. To evaluate the performance of our model, we conduct simulation studies to assess parameter estimation accuracy. Our results demonstrate that the proposed approach outperforms existing methods reported in the literature. Finally, we apply our model to E. coli microarray data, illustrating its practical utility in gene expression analysis.

28. Network Bootstrap Using Overlapping Partitions

[03.M2.I54, (page 29)]

Sayan CHAKRABARTY, *University of Michigan*
Sayan CHAKRABARTY, *University of Michigan*
Elizaveta LEVINA, *University of Michigan*

Bootstrapping network data efficiently is a challenging task. The existing methods tend to make strong assumptions on both the network structure and the statistics being bootstrapped, and are computationally costly. This paper introduces a general algorithm, OPBoot for network bootstrap that partitions the network into multiple overlapping subnetworks and then aggregates results from bootstrapping these subnetworks to generate a bootstrap sample of the network statistic of interest. This approach tends to be much faster than competing methods as most of the computations are done on smaller subnetworks. We show that OPBoot is consistent in distribution for a large class of network statistics under minimal assumptions on the network structure, and demonstrate with extensive numerical examples that the bootstrap confidence intervals produced by OPBoot attain good coverage without substantially increasing interval lengths in a fraction of the time needed for running competing methods.

29. The Decaying MAR Framework: Model Doubly Robust Causal Inference with Partially Labeled Data

[02.M1.I22, (page 16)]

Abhishek CHAKRABORTTY, *Texas A&M University*

In modern large-scale observational studies, data collection constraints often result in partially labeled datasets, posing challenges for reliable causal inference, especially due to potential labeling bias and

relatively small size of the labeled data. This paper introduces a decaying missing-at-random (decaying MAR) framework and associated approaches for doubly robust causal inference on treatment effects in such semi-supervised (SS) settings. This simultaneously addresses selection bias in the labeling mechanism and the extreme imbalance between labeled and unlabeled groups, bridging the gap between the standard SS and missing data literatures, while throughout allowing for confounded treatment assignment and high-dimensional confounders under appropriate sparsity conditions. To ensure robust causal conclusions, we propose a bias-reduced SS (BRSS) estimator for the average treatment effect, a type of 'model doubly robust' estimator appropriate for such settings, establishing asymptotic normality at the appropriate rate under decaying labeling propensity scores, provided that at least one nuisance model is correctly specified. Our approach also relaxes sparsity conditions beyond those required in existing methods, including standard supervised approaches. Recognizing the asymmetry between labeling and treatment mechanisms, we further introduce a de-coupled BRSS (DC-BRSS) estimator, which integrates inverse probability weighting (IPW) with bias-reducing techniques in nuisance estimation. This refinement further weakens model specification and sparsity requirements. Numerical experiments confirm the effectiveness and adaptability of our estimators in addressing labeling bias and model misspecification.

30 . Minimax And Adaptive Transfer Learning for Nonparametric Classification under Distributed Differential Privacy Constraints

[01.M2.I2, (page 4)]

Abhinav CHAKRABORTY, *Columbia University*
Arnab AUDDY, *Ohio State University*
Tony CAI, *University of Pennsylvania*

We study minimax and adaptive transfer learning for nonparametric classification under a posterior drift model with distributed differential privacy. Our heterogeneous framework allows for varying sample sizes, privacy levels, and data distributions across servers. We characterize the minimax misclassification rate and reveal trade-offs between privacy and accuracy, including phase transitions. We also propose a data-driven adaptive classifier that nearly attains the optimal rate while respecting privacy constraints. Simulations and real data validate our theoretical findings.

31. Robust probabilistic inference via a constrained transport metric

[03.M1.I51, (page 28)]

Abhisek CHAKRABORTY, *Eli Lilly and Company*
Abhisek CHAKRABORTY, *Eli Lilly and Company*
Anirban BHATTACHARYA, *Texas A&M University*
Debdeep PATI, *University of Wisconsin, Madison*

Flexible Bayesian models are typically constructed using limits of large parametric models with a multitude of parameters that are often difficult to interpret. In this article, we offer a novel alternative by constructing an exponentially tilted empirical likelihood carefully designed 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 robustified 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 the presence of outliers. Our proposed transport metric enjoys great computational simplicity and is inherently parallelizable, exploiting the Sinkhorn regularization for discrete optimal transport problems. We demonstrate superior performance of our methodology when compared against the state-of-the-art robust Bayesian inference methods. We also demonstrate the equivalence of our approach with a non-parametric Bayesian formulation under a suitable asymptotic framework, thereby testifying to its flexibility.

32. TBD

[02.M2.I30, (page 19)]

Antik CHAKRABORTY, *Purdue University*

TBA

33 . Statistical Inference for Subgraph Densities Under Random Sampling from Network Data

[03.M2.I54, (page 29)]

Nilanjan CHAKRABORTY, *Missouri University of Science and Technology*
Ayoushman BHATTACHARYA, *Washington University in Saint Louis*
Soumen LAHIRI, *Washington University in Saint Louis*

In this talk, we discuss the framework for obtaining the statistical guarantees for subgraph densities of a population network under without replacement sampling (SRSWOR). Under this sampling scheme, we establish the asymptotic normality of the Horwitz-Thompson (HT) estimator for the population sub-

graph densities under minimal assumptions. We also establish the joint asymptotic normality of two subgraph densities which is crucial in establishing weak convergence of the global transitivity of the sampled graph. To facilitate the inferential procedures, we provide a jackknife and a bootstrap estimator of the unknown population variance and establish its consistency. Our results find a useful application to the problem of testing the equality of two population graphs using the subgraph densities as the test statistic. Finally, we present a simulation study and a real data analysis that corroborate our theoretical findings.

(This is a joint work with Ayoushman Bhattacharya and Prof. Soumen Lahiri.

34. On the Statistical Properties of Generative Adversarial Models for Low Intrinsic Data Dimension

[03.M2.I56, (page 30)]

Saptarshi CHAKRABORTY, *University of Michigan*
Peter BARTLETT,

Despite the remarkable empirical successes of Generative Adversarial Networks (GANs), the theoretical guarantees for their statistical accuracy remain rather pessimistic. In particular, the data distributions on which GANs are applied, such as natural images, are often hypothesized to have an intrinsic low-dimensional structure in a typically high-dimensional feature space, but this is often not reflected in the derived rates in the state-of-the-art analyses. In this paper, we attempt to bridge the gap between the theory and practice of GANs and their bidirectional variant, Bi-directional GANs (BiGANs), by deriving statistical guarantees on the estimated densities in terms of the intrinsic dimension of the data and the latent space. We analytically show that if one has access to n samples from the unknown target distribution and the network architectures are properly chosen, the expected Wasserstein-1 distance of the estimates from the target scales as $\mathcal{O}\left(n^{-1/d\_\mu}\right)$ for GANs and $\tilde{\mathcal{O}}\left(n^{-1/(d\_\mu+\ell)}\right)$ for BiGANs, where d_μ and ℓ are the upper Wasserstein-1 dimension of the data-distribution and latent-space dimension, respectively. The theoretical analyses not only suggest that these methods successfully avoid the curse of dimensionality, in the sense that the exponent of n in the error rates does not depend on the data dimension but also serve to bridge the gap between the theoretical analyses of GANs and the known sharp rates from optimal transport literature. Additionally, we demonstrate that GANs can effectively achieve the minimax optimal rate even for non-smooth underlying distributions, with the use of interpolating generator networks.

35 . Flexible Empirical Bayesian Approaches to Pharmacovigilance for Simultaneous Signal Detection and Signal Strength Estimation in Spontaneous Reporting Systems Data

[04.M2.I74, (page 38)]

Saptarshi CHAKRABORTY, *University at Buffalo*
Yihao TAN, *University at Buffalo*
Marianthi MARKATOU, *University at Buffalo*

Inferring adverse events (AEs) of medical products from Spontaneous Reporting Systems (SRS) databases is a core challenge in contemporary pharmacovigilance. Bayesian methods for pharmacovigilance are attractive for their rigorous ability to simultaneously detect potential AE signals and estimate their strengths/degrees of relevance. However, existing Bayesian and empirical Bayesian methods impose restrictive parametric assumptions and/or demand substantial computational resources, limiting their practical utility. This paper introduces a suite of novel, scalable empirical Bayes methods for pharmacovigilance that utilize flexible non-parametric priors and custom, efficient data-driven estimation techniques to enhance signal detection and signal strength estimation at a low computational cost. Our highly flexible methods accommodate a broader range of data and achieve signal detection performance comparable to or better than existing Bayesian and empirical Bayesian approaches. More importantly, they provide coherent and high-fidelity estimation and uncertainty quantification for potential AE signal strengths, offering deeper insights into the comparative importance and relevance of AEs. Extensive simulation experiments across diverse data-generating scenarios demonstrate the superiority of our methods in terms of accurate signal strength estimation, as measured by replication root mean squared errors. Additionally, our methods maintain or exceed the signal detection performance of state-of-the-art techniques, as evaluated by frequentist false discovery rates and sensitivity metrics. Applications on FDA FAERS data for the statin group of drugs reveal interesting insights through Bayesian posterior probabilities.

36 . An Empirical Evaluation of GTI Rankings Through Event Data and Media Framing

[03.A1.C2, (page 33)]

Sagnik CHAKRAVARTY, *University of Maryland*

College Park

The Global Terrorism Index (GTI), developed by the Institute for Economics and Peace, is widely used to rank countries based on the impact of terrorism, yet its methodological opacity, exclusion of state-led violence, and overreliance on selective definitions raise serious concerns about validity and fairness. This project aims to interrogate the GTI by building a supervised machine learning model that aligns event-based terrorism data with narrative framings in global media coverage. The primary dataset will be the Global Terrorism Database(GTD) maintained by the University of Maryland, which includes over 200,000 records of terrorist events with rich metadata such as actor type, location, casualties, and narrative summaries. To address the GTI's blind spots, additional conflict data will be sourced from ACLED (Armed Conflict Location & Event Data) and UNDP human security metrics, allowing us to integrate protest suppression, state-perpetrated violence, and broader patterns of insecurity that often go uncounted. In parallel, I will scrape and compile a corpus of news articles using Playwright (Python) from a diverse set of international and regional sources, such as The Hindu, Times of India, Al Jazeera, Reuters, BBC, CNN, NDTV, and Russia Today, focusing on the period between 2021 and 2025. Articles will be filtered using keywords like "terrorist," "insurgent," "freedom fighter," "militant," and "state crackdown," and metadata including publication source, date, and content will be saved. Each article will then be hand-coded for sentiment toward actors and states involved, enabling the training of a supervised learning model to predict whether a violent event—based on both factual details and framing tone—would be considered terrorism under GTI's criteria. The model will help reveal systematic framing biases and regional inconsistencies, and inform the creation of a bias-aware terrorism score that accounts for overlooked or misclassified events. Case studies such as India—ranked unusually high in the GTI despite relatively stable conflict activity—will be used to demonstrate how framing and selection may distort global terrorism perceptions. The final output of the project will be an alternative scoring framework grounded in transparency, human-coded sentiment, and machine learning, contributing a critical lens to how terrorism is quantified in global discourse.

37. Point Prediction of Streaming Data

[01.A1.I11, (page 11)]

Aleena CHANDA, *University of Nebraska-Lincoln*
Bertrand CLARKE, *University of Nebraska-Lincoln*

We present two new approaches for point prediction with streaming data based on a) the Count-Min sketch and b) Gaussian Process Priors with random bias. The methods are intended for the most general case where no true model can be usefully formulated for the data stream. In statistical contexts, this is often called the $\mathcal{M}$-open problem class. For the Count Min Sketch method we show that the predicted distribution function $\hat{F}$ converges to F under the assumption that the data consists of i.i.d samples from a fixed distribution function F. To implement the Gaussian Process Prior methods, we used representative subsets based on streaming $K-$ means to keep the dimension of the variance matrices bounded.

We form four versions of our hash function based predictor (HBP) and compare them to six other predictors that are based on existing methods. Four are fully Bayesian and the other two are derived from the Shtarkov solution. For comparisons we use real data. In our experiments, we use absolute cumulative error as criterion for predictive success. Preliminary experiments suggest that the one-pass median version of our method performs the best compared to any other methods for data whose complexity is not based on spread. For spread-complex data, often Gaussian Processes are best. In some cases when our method is best, some of the Shtarkov methods are comparable. We argue that when this happens, our method is preferred over Shtarkov methods because of its simplicity.

38. Bayesian Semiparametric Functional Mixed Effects Drift-Diffusion Models for Cognitive Control Leveraging Stop-signal Tasks

[04.M2.I72, (page 37)]

Noirrit Kiran CHANDRA, *The University of Texas at Dallas*
Farabi Raihan SHUVO, *The University of Texas at Dallas*
Stacie WARREN, *The University of Texas at Dallas*

Cognitive control refers to a set of dynamic processes in the brain that encompasses stimulus-driven response mechanisms, goal-oriented decision-making, and the capacity to learn and adapt based on past decisions and their outcomes. Understanding the developmental progression of cognitive control in children and youth is crucial for identifying critical growth periods and ensuring optimal developmental outcomes. The Stop-signal task (SST) is a widely used paradigm to assess cognitive control in humans, where participants engage in a series of tasks and their behavioral

metrics are recorded. Wiener drift-diffusion models (DDMs) are widely used to analyze behavioral data from SSTs as several cognitive psychology studies validated the interpretability of DDM parameters with respect to task conditions. We propose a comprehensive class of Bayesian semiparametric functional mixed drift-diffusion models for the behavioral metrics of SST. Cognitive control is inherently dynamic, unfolding across successive trials during task performance. Our proposed advancements in DDMs aim to incorporate these insights by accounting for the sustained and transient natures of proactive and reactive controls, respectively, incorporating performance monitoring as quantifiable functions, reflecting the dynamic interplay between the control mechanisms, and considering across-subject variability via nonparametric random effects. Our approach provides insights into how cognitive control evolves during task performance and how external factors influence these processes—critical aspects often overlooked in traditional cognitive research.

39 . Performance Guaranteed Confidence Sets of Ranks

[Student Poster Competition, (page 6)]

Onrina CHANDRA, *Rutgers University*
Minge XIE,

Ranks of populations are often based on estimates of certain latent features of the population. Due to sample randomness, it is of interest to quantify the uncertainty associated with the estimated ranks. This task is especially important in the situations in which the latent features are not well separated among some of the populations resulting in a nonignorable portion of wrongly ordered estimated ranks. Uncertainty quantification can help mitigate some of the issues and give us a fuller picture. However, this is very challenging because the ranks are discrete parameters and the standard inference methods developed under regularity conditions do not apply. Bayesian methods are sensitive to prior choices while large sample-based methods do not work since the central limit theorem fail to hold for the estimated ranks. In this article, we propose a Repro Samples Method to address this nontrivial irregular inference problem by developing a confidence set for the true unobserved ranks of the populations. The confidence set obtained has finite sample coverage guarantee and the method can handle difficult near-tie cases. The effectiveness of the proposed development is illustrated using simulation studies and real data examples of ranking the performance of US VHA facilities in their service to their diabetic patients, Men's basketball teams of the 2023-24 Big 10 Conference, teams playing in English Premier League 2024 and the Forbes Billionaires 2024 dataset.

40. Inference for projection parameters in Linear Regression

[Student Poster Competition, (page 6)]

Woonyoung CHANG, *Carnegie Mellon University*
Arun KUCHIBHOTLA, *Carnegie Mellon University*
Alessandro RINALDO, *University of Texas at Austin*

We study statistical inference in a misspecified linear regression model with a stochastic design where the number of covariates d increases with the sample size n. This problem has been studied under a variety of assumptions in the literature. When $d = o(n^{1/2})$, it is known that the traditional Wald confidence intervals based on the asymptotic normality of the least squares estimator and the sandwich variance estimator are asymptotically valid. In this work, we develop a bias correction that ensures $\sqrt{n}$-consistency and asymptotically normality of the resulting estimator for linear contrasts of the projection parameter under weak moment conditions, extending validity to $d = o(n^{2/3})$ with an explicit bound on the rate of convergence to normality. Additionally, we establish the consistency of the sandwich variance estimator under the same dimension scaling, ensuring valid Wald inference. We also leverage recent inference methods that do not require variance estimation, leading to sharper miscoverage rate guarantees. The proposed methodology is assumption-lean and avoids strong model, distributional, or sparsity assumptions, making it broadly applicable in empirical settings.

41 . Revisiting Total Variation Denoising: New Perspectives and Generalizations

[02.A1.I34, (page 21)]

Sabyasachi CHATTERJEE, *University of Illinois at Urbana Champaign*

Total Variation Denoising (TVD) is a fundamental denoising/smoothing method. Here, we identify a new local minmax/maxmin formula producing two estimators which sandwich the univariate TVD estimator at every point. Operationally, this formula gives a local definition of TVD as a minmax/maxmin of a simple function of local averages. Moreover we find that this minmax/maxmin formula is generalizeable and can be used to define other TVD like estimators. We further propose and study higher order polynomial versions of TVD which are defined pointwise lying between minmax and maxmin optimizations of

penalized local polynomial regressions over intervals of different scales. These appear to be new nonparametric regression methods, different from usual Trend Filtering and any other existing method in the nonparametric regression toolbox. We call these estimators Minmax Trend Filtering (MTF). We show how the proposed local definition of TVD/MTF estimator makes it tractable to bound pointwise estimation errors in terms of a local bias variance like tradeoff. This type of local analysis of TVD/MTF is new and arguably simpler than existing analyses of TVD/Trend Filtering. In particular, apart from minimax rate optimality over bounded variation and piecewise polynomial classes, our pointwise estimation error bounds also enable us to derive local rates of convergence for (locally) Holder Smooth signals. These local rates offer a new pointwise explanation of local adaptivity of TVD/MTF instead of global (MSE) based justifications.

42. PriME: Privacy-aware Membership profile Estimation in networks

[Student Poster Competition, (page 6)]

Sayak CHATTERJEE, *University of Pennsylvania*
Abhinav CHAKRABORTY, *University of Pennsylvania*
Sagnik NANDY, *University of Chicago*

This paper presents a novel approach to estimating community membership probabilities for network vertices generated by the Degree Corrected Mixed Membership Stochastic Block Model while preserving individual edge privacy. Operating within the ε-edge local differential privacy framework, we introduce an optimal private algorithm based on a symmetric edge flip mechanism and spectral clustering for accurate estimation of vertex community memberships. We conduct a comprehensive analysis of the estimation risk and establish the optimality of our procedure by providing matching lower bounds to the minimax risk under privacy constraints. To validate our approach, we demonstrate its performance through numerical simulations and its practical application to real-world data. This work represents a significant step forward in balancing accurate community membership estimation with stringent privacy preservation in network data analysis.

43. The Effect of Restrictive Interactions between Susceptible and Infected Individuals on the Prognosis of an Epidemic

[01.E1.I16, (page 12)]

Shirshendu CHATTERJEE, *City University of New York*

We will discuss some adaptations of the standard epidemic models to incorporate various kinds of restrictions on the interaction between susceptible and infected individuals and study the effect of such restrictions on the prognosis of an epidemic for those models. In one case, we study the effect of avoiding known infected neighbors on the persistence of a recurring infection process. In another case, we study the effect of isolation protocols for infected individuals and scenarios in a vigilant society.

44. TBD

[01.A1.I11, (page 11)]

Snigdhansu CHATTERJEE, *University of Maryland, Baltimore County*

TBA

45. Neural networks can learn any low complexity pattern

[Bahadur Memorial Lecture, (page 6)]

Sourav CHATTERJEE, *Stanford University*
Timothy SUDIJONO,

I will present recent work showing that feedforward neural networks can, in principle, learn patterns that can be expressed as a short program. An example is as follows. Let N be a large number, and suppose our data consists of a sample of X's and Y's, where each X is a randomly chosen number between 1 and N, and the corresponding Y is 1 if X is a prime and 0 if not. The sample size n is negligible compared to N. If we fit a neural network to this data which is "sparsest" in a suitable sense, it turns out that the network will be able to accurately predict if a newly chosen X is a prime or not. This is because the property of being prime can be tested by a short program. This is based on joint work with Tim Sudijono. The talk will be accessible to those with no background in neural networks.

46. Neymanian inference in randomized experiments

[02.A1.I32, (page 20)]

Ambarish CHATTOPADHYAY, *Stanford University*
Guido IMBENS, *Stanford University*

In his seminal 1923 work, Neyman studied the variance estimation problem for the difference-in-means estimator of the average treatment effect in completely randomized experiments. He proposed a variance estimator that is conservative in general and unbiased under homogeneous treatment effects. While widely used under complete randomization, there is no unique or natural way to extend this esti-

mator to more complex designs. To this end, we show that Neyman's estimator can be alternatively derived in two ways, leading to two novel variance estimation approaches: the imputation approach and the contrast approach. While both approaches recover Neyman's estimator under complete randomization, they yield fundamentally different variance estimators for more general designs. In the imputation approach, the variance is expressed in terms of observed and missing potential outcomes and then estimated by imputing the missing potential outcomes, akin to Fisherian inference. In the contrast approach, the variance is expressed in terms of unobservable contrasts of potential outcomes and then estimated by exchanging each unobservable contrast with an observable contrast. We examine the theoretical properties of both approaches, showing that for a large class of designs, each produces non-negative, conservative variance estimators that are unbiased in finite samples or asymptotically under homogeneous treatment effects.

47 . Leveraging the blessing of dimensions for scalable Bayesian inference on covariance matrices

[03.M1.I49, (page 27)]

Shounak CHATTOPADHYAY, *University of California, Los Angeles*
Anru ZHANG, *Duke University*
David DUNSON, *Duke University*

Bayesian factor analysis provides an elegant approach for probabilistic inference on high-dimensional covariance matrices, decomposing the covariance as a sum of a low-rank and diagonal matrix. Posterior computation in such models typically involves Markov chain Monte Carlo (MCMC) algorithms such as Gibbs sampling steps that alternately update the latent factors, factor loadings, and residual variances. In this talk, we exploit a remarkable blessing of dimensionality phenomenon to develop a provably accurate approximate-posterior for the covariance matrix, completely bypassing the need for MCMC via an embarrassingly parallel computational framework. The proposed Factor Analysis with BLEssing of dimensionality (FABLE) approach relies on a first-stage singular value decomposition to estimate the latent factors, then defines a jointly conjugate prior for the loadings and residual variances. The accuracy of the resulting approximate-posterior for the covariance improves with increasing dimensionality. FABLE has excellent performance in high-dimensional covariance matrix estimation, including producing well-calibrated credible intervals, both theoretically and through simulation experiments. We also demonstrate the strength of our approach in terms of accurate inference and computational efficiency by applying it to a gene expression data set.

48 . Consistent DAG selection for Bayesian causal discovery under general error distributions

[02.M2.I30, (page 19)]

Anamitra CHAUDHURI, *Texas A&M University*
Anirban BHATTACHARYA, *Texas A&M University*
Yang NI, *Texas A&M University*

We consider the problem of learning the underlying causal structure among a set of variables, which are assumed to follow a Bayesian network or, more specifically, a linear acyclic structural equation model (SEM) with the associated errors being independent and allowed to be non-Gaussian. A Bayesian hierarchical model is proposed with the objective of identifying the true data-generating directed acyclic graph (DAG) structure where the nodes and edges represent the variables and the direct causal effects, respectively. Moreover, incorporating the information of non-Gaussian errors, we characterize the distribution equivalence class of the true DAG, which specifies the best possible extent up to which the DAG can be identified based on purely observational data. Furthermore, under the consideration that the errors are distributed as some scale mixture of Gaussian, where the mixing distribution is unspecified, and mild distributional assumptions, we establish that the posterior probability of the distribution equivalence class of the true DAG converges to unity as the sample size grows. This shows that the proposed method achieves the posterior DAG selection consistency. Simulation studies are presented to illustrate the results, where we also demonstrate different rates of divergence of the associated Bayes factors varying over the competing DAGs.

49 . Bayesian Mechanistic Model of Spatio-Temporal Dynamics in Invasive Tree

[Student Poster Competition, (page 6)]

Jiaqi CHEN, *University of Nebraska-Lincoln*
Yawen GUAN,
Huijing DU,

Modeling spatio-temporal data is a common task, and it is crucial to capture the underlying mechanics driving the observed dynamics. Invasive tree species, such as the eastern redcedar, exhibit complex behaviors characterized by spatial diffusion and localized growth, influenced by environmental factors like

proximity to water sources. Partial differential equations (PDEs) provide an effective framework for incorporating these mechanics through physical principles. We propose a Bayesian mechanistic framework based on a diffusion-growth PDE to model the spatio-temporal dynamics of invasive trees and their relationship with environmental factors. To address the computational challenges of solving PDEs, we employ numerical solutions combined with homogenization techniques, enabling scalable inference. The proposed method is validated using simulated data and applied to satellite-derived tree cover data from the Sandhills region of Nebraska to infer dynamics of tree encroachment.

50. Agnostic Characterization of Interference in Randomized Experiments

[02.M1.I22, (page 16)]

David CHOI, *Carnegie Mellon University*

We give an approach for characterizing interference by lower bounding the number of units whose outcome depends on selected groups of treated individuals, such as depending on the treatment of others, or others who are at least a certain distance away. The approach is applicable to randomized experiments with binary-valued outcomes. Asymptotically conservative point estimates and one-sided confidence intervals may be constructed with no assumptions beyond the known randomization design, allowing the approach to be used when interference is poorly understood, or when an observed network might only be a crude proxy for the underlying social mechanisms. Point estimates are equal to Hajek-weighted comparisons of units with differing levels of treatment exposure. Empirically, we find that the width of our interval estimates is competitive with (and often smaller than) those of the EATE, an assumption-lean treatment effect, suggesting that the proposed estimands may be intrinsically easier to estimate than treatment effects.

51. A Graph-based Approach to Estimating the Number of Clusters in High-dimensional Settings

[03.E1.I64, (page 34)]

Lynna CHU, *Iowa State University*
Yichuan BAI, *Iowa State University*
Lynna CHU, *Iowa State University*

We consider the problem of estimating the number of clusters (k) in a dataset. We propose a nonparametric approach to the problem that utilizes similarity graphs to construct a robust statistic that effectively captures similarity information among observations. This graph-based statistic is applicable to datasets of any dimension, is computationally efficient to obtain, and can be paired with any clustering technique. Asymptotic theory is developed to establish the selection consistency of the proposed approach. Simulation studies demonstrate that the graph-based statistic outperforms existing methods for estimating k, especially in the high-dimensional setting. We illustrate its utility on an RNA-seq dataset.

52. Stability and Shrinkage Selection in High Dimensional Logistic Regression

[03.E1.I63, (page 33)]

Jennifer CLARKE, *University of Nebraska-Lincoln*
Bertrand CLARKE, *University of Nebraska-Lincoln*
Caleb HOLMBECK, *University of Nebraska-Lincoln*
Laura KRESTY, *University of Michigan*

In regression contexts where p » n it is common to invoke a shrinkage (or regularization) technique that determines the appropriate model by minimizing a penalized empirical risk criterion. We explore the stability of the predictive error as a metric for comparing and selecting a regularization method in high dimensional logistic regression, with evolutionary computation and subsampling of representative variable sets for optimization.

53. Marginal Structural Modeling for Causal Inference of E-Cigarette Vaping and Smoking Relapse – Six-Wave Longitudinal Study 2013 - 2021

[01.E1.I13, (page 12)]

Daisy DAI, *UNMC*

Marginal Structural Models (MSMs) offer potential for causal inference for former smokers who vape e-cigarettes after quitting cigarettes, addressing complex time-varying confounders and prior exposures. Six longitudinal waves of the Population Assessment of Tobacco and Health (PATH) were analyzed, starting with former cigarette smokers aged >18 who were recent (quit <12 months) or long-term (quit >12 months) quitters at baseline. These individuals were followed up in waves 2-6 (2013-2021, with sample sizes of 4,919 individuals and 22,030 observations. Across all waves, the prevalence of smoking relapse was 12%, with higher relapse rates among current e-cigarette users (25% vs. 11% for non-users) and those who vaped more frequently (some days: 39%, every day: 21%, vs. none: 11%). Among recent quitters (n=1090), e-cigarette use was not associated with smoking relapse at follow-up after covariate adjustment. In long-term cigarette quit-

ters (n=3782), both some-day (AOR[95%CI]=5.44 [2.34, 12.65], p<0.0001) and everyday e-cigarette users (AOR[95%CI]=4.13 [1.9, 8.95], p=0.0003) had higher odds of subsequent smoking relapse as compared to non-e-cigarette users after covariate adjustment. Stratified analysis reveals that Hispanic long-term quitters face an elevated risk of smoking relapse linked to vaping (AOR[95%CI]=10.41 [3.34, 32.43], p<0.0001). Marginal structural modeling of six-wave longitudinal data offers causal insights into the relationship between vaping and smoking relapse among long-term quitters, providing evidence to inform tobacco regulations on e-cigarettes as a harm reduction alternative in consideration of the potential risk of smoking relapse.

54. A hybrid mixture of factor analyzers approach for characterizing high dimensional data

[02.A1.I37, (page 22)]

Fan DAI, *Michigan Technological University*
Kazeem KAREEM, *Michigan Technological University*

Clustering high-dimensional data is of the major interest in modern scientific research, where the data complexity and dimensionality often cause difficulties in the implementation of traditional statistical techniques. We develop a hybrid mixture of factor analyzers approach to explain the variability of grouped data with a large set of features using a few latent factors. The proposed model allows for varying numbers of factors for different groups and adopts a matrix-free computational framework for efficient parameter estimation. Our method is applied to cluster Wisconsin breast cancer data, recognize USPS digit images, and characterize lymphoma gene expression data.

55. Hypothesis selection via sample splitting for valid powerful testing in matched observational studies

[Student Paper Competition 2, (page 5)]

Abhinandan DALAL, *University of Pennsylvania*
William BEKERMAN, *University of Pennsylvania*
Carlo DEL NINNO, *Formerly of the World Bank*
Dylan SMALL, *University of pennsylvania*

Observational studies are valuable tools for inferring causal effects in the absence of controlled experiments. However, these studies may be biased due to the presence of some relevant, unmeasured set of covariates. One approach to mitigate this concern is to identify hypotheses likely to be more resilient to hidden biases by splitting the data into a planning sample for designing the study and an analysis sample for making inferences. We devise a flexible method for selecting hypotheses in the planning sample when an unknown number of outcomes are affected by the treatment, allowing researchers to gain the benefits of exploratory analysis and still conduct powerful inference under concerns of unmeasured confounding. We investigate the theoretical properties of our method and conduct extensive simulations that demonstrate pronounced benefits, especially at higher levels of allowance for unmeasured confounding. Finally, we demonstrate our method in an observational study of the multi-dimensional impacts of a devastating flood in Bangladesh.

56. B-MASTER: Scalable Bayesian multivariate regression analysis for selecting targeted essential regressors to identify the key genera in microbiome-metabolite relation dynamics

[01.A1.I10, (page 11)]

Priyam DAS, *Virginia Commonwealth University*

TBA

57. Scalable Efficient Inference in Complex Surveys through Targeted Resampling of Weights

[Student Poster Competition, (page 7)]

Snigdha DAS, *Texas A&M University*
Dipankar BANDYOPADHYAY, *Virginia Commonwealth University*
Debdeep PATI, *University of Wisconsin - Madison*

Survey data often arises from complex sampling designs, such as stratified or multistage sampling, with unequal inclusion probabilities. When sampling is informative, traditional inference methods yield biased estimators and poor coverage. Classical pseudo-likelihood based methods provide accurate asymptotic inference but lack finite-sample uncertainty quantification and the ability to integrate prior information. Existing Bayesian approaches, like the Bayesian pseudo-posterior estimator and weighted Bayesian bootstrap, have limitations; the former struggles with uncertainty quantification, while the latter is computationally intensive and sensitive to bootstrap replicates. To address these challenges, we propose the Survey-adjusted Weighted Likelihood Bootstrap (S-WLB), which resamples weights from a carefully chosen distribution centered around the underlying sampling weights. S-WLB is computationally efficient, theoretically consistent, and delivers finite-sample uncertainty intervals which are proven to be asymptotically valid. We demonstrate its performance through simulations and applications to na-

tionally representative survey datasets like NHANES and NSDUH.

58. Differentially Private Bayesian Tests

[Student Poster Competition, (page 7)]
Saptati DATTA, *Texas A&M University*
Abhisek CHAKRABORTY,

Differential privacy has emerged as a significant cornerstone in the realm of scientific hypothesis testing utilizing confidential data. When data are not confidential, Bayesian tests are widely used in reporting scientific discoveries, as they effectively address the key criticisms of p-values, namely lack of interpretability and inability to quantify evidence in support of competing hypotheses. In this article, we introduce a novel framework for differentially private Bayesian hypothesis testing, thereby expanding the applicability of Bayesian testing to confidential data. This framework naturally arises from a principled data-generative mechanism, ensuring that the resulting inferences retain interpretability while maintaining privacy. Further, by focusing on differentially private Bayes factors based on test statistics, we circumvent the need to model the complete data generative mechanism and ensure substantial computational benefits. We also provide a set of sufficient conditions to establish Bayes factor consistency under the proposed framework. Finally, the utility of the proposed methodology is showcased via several numerical experiments.

59. "Unraveling Disease Mysteries: Statistical Models Reveal Cellular Conversations using Spatial Transcriptomics data."

[Special Invited Session 3, (page 20)]
Susmita DATTA, *University of Florida*
Dongyuan WU, *Moderna Inc. and University of Florida*

Understanding cell microenvironments from spatially resolved transcriptomics data is a cutting-edge approach in biomedical research. This innovative method enables scientists to investigate the spatial organization of cells near diseased tissues and identify their inter - and intracellular communications through biochemical signaling, crucial for elucidating disease mechanisms and developing targeted treatments. Traditionally, most computational methods provide ad hoc measurements to estimate intercellular communication. While straightforward, these methods often lack the accuracy and reliability that robust statistical models can offer. To address these limitations, our research proposes a novel generalized linear regression model known as Bayesian Tweedie Modeling of Communications (BATCOM). This model is designed to infer cellular communications from spatially resolved transcriptomics data, particularly spot-based data, by estimating communication scores between cell types while considering their spatial distances. BATCOM offers a nuanced and statistically sound approach to understanding cellular interactions. By incorporating spatial distance into the communication score estimations, BATCOM provides a more accurate representation of how cells interact within their microenvironments, significantly improving upon traditional methods that often overlook the spatial aspect of cellular communications. Additionally, we explore a frequentist approach using the generalized additive model (GAM) framework. Implemented in the associated TWCOM software in R, this approach enhances scalability and integration, making it more user-friendly for researchers. We demonstrate the superiority of our method using single-cell and spatial RNA-seq data for cutaneous squamous cell carcinoma, the second most common skin cancer in the USA. These advancements in statistical modeling are crucial for advancing our understanding of disease mechanisms. Accurate inference of cellular interactions can reveal new insights into how diseases develop and progress at the cellular level, informing the development of more effective treatments and interventions. By integrating BATCOM and TWCOM into user-friendly software, we ensure these advanced statistical methods are accessible to a wide range of researchers, accelerating biomedical discoveries and improving patient outcomes. Our research enhances our understanding of disease mechanisms, paving the way for new discoveries and therapeutic strategies in biomedical research.

60 . Spatially Varying Gene Regulatory Networks via Bayesian Nonparametric Covariate-Dependent Directed Cyclic Graphical Models

[Student Poster Competition, (page 7)]
Trisha DAWN, *Texas A & M University*
Yang NI,

Single-cell RNA sequencing (scRNA-seq) has advanced biological research but lacks spatial context. Spatial transcriptomics addresses this by bridging gene expression to spatial location, enabling insights into spatially regulated functions. A key challenge is inferring spatially varying gene regulatory networks (svGRNs), which capture region-specific gene interactions, unlike traditional GRNs. While prior

work focuses on learning undirected or acyclic graphs, these fail to capture feedback loops in such a context. We introduce Bayesian nonparametric directed cyclic graphs with covariates (BNP-DCGx), a novel method for inferring directed cyclic svGRNs from spatial transcriptomics data. Our model includes a covariate-dependent random partition layer within a Bayesian hierarchical framework, enabling inference of spatially varying DCGs. We develop a parallel-tempered MCMC sampler for posterior inference with stability guarantees required for DCGs. We validate BNP-DCGx through simulations and analyze human DLPFC data, revealing spatially dynamic regulatory interactions. This is, to our knowledge, the first method to model svGRNs with feedback loops while capturing spatial heterogeneity.

61. Centile Curve Modelling for Football Performance in Athletic and General Populations

[03.A1.C2, (page 33)]

Praveen D CHOUGALE, *Indian Institute of Technology Bombay*
Praveen D CHOUGALE, *Indian Institute of Technology Bombay*
Prof. Usha ANANTHAKUMAR, *Shailesh J. Mehta School of Management, I.I.T.Bombay, Powai, Mumbai*

Performance benchmarking in football often lacks standardized, population-specific reference data—particularly in the Indian context. This study addresses that gap by developing a statistical framework for centile-based performance evaluation of male and female football players, compared against the general population. Standardizing such benchmarks is crucial for talent identification, training personalization, and athlete health monitoring. Using Generalized Additive Models for Location, Scale, and Shape (GAMLSS), sex-specific centile curves were constructed for key physical metrics, including jump height, flight time, and reactive strength index. GAMLSS enables flexible modelling of skewed and age-dependent performance data. Models were selected using Akaike Information Criterion (AIC) and validated through mean absolute error (MAE) across percentiles and age groups. While percentile differences between athletes and the general population were modest at lower levels, they widened considerably at the higher end, indicating performance specialization in football athletes. The analysis also revealed distinct sex-based performance trends—males generally exhibited higher power outputs, whereas female athletes showed broader variability, possibly reflecting diverse training exposure and physiological factors. These benchmarks offer actionable insights for coaches, trainers, and sports scientists.

62. No-Regret Generative Modeling via Parabolic Monge-Ampère PDE

[01.M2.I3, (page 4)]

Nabarun DEB, *University of Chicago*
Tengyuan LIANG, *University of Chicago*

We introduce a novel generative modeling framework based on a discretized parabolic Monge-Ampère PDE, which emerges as a continuous limit of the Sinkhorn algorithm commonly used in optimal transport. Our method performs iterative refinement in the space of Brenier maps using a mirror gradient descent step. We establish theoretical guarantees for generative modeling through the lens of no-regret analysis, demonstrating that the iterates converge to the optimal Brenier map under a variety of step-size schedules. As a technical contribution, we derive a new Evolution Variational Inequality tailored to the parabolic Monge-Ampère PDE, connecting geometry, transportation cost, and regret. Our framework accommodates non-log-concave target distributions, constructs an optimal sampling process via the Brenier map, and integrates favorable learning techniques from generative adversarial networks and score-based diffusion models. As direct applications, we illustrate how our theory paves new pathways for generative modeling and variational inference.

63. Estimating sparse direct effects in multivariate regression with the spike-and-slab LASSO

[03.M2.I55, (page 29)]

Sameer DESHPANDE, *sameer.deshpande@wisc.edu*
Shen YUNYI, *MIT*
Claudia SOLÍS-LEMUS, *University of Wisconsin–Madison*

The multivariate regression interpretation of the Gaussian chain graph model simultaneously parametrizes (i) the direct effects of p predictors on q outcomes and (ii) the residual partial covariances between pairs of outcomes. We introduce a new method for fitting sparse versions of these models with spike-and-slab LASSO (SSL) priors. We develop an Expectation Conditional Maximization algorithm to obtain sparse estimates of the $p \times q$ matrix of direct effects and the $q \times q$ residual precision matrix. Our algorithm iteratively solves a sequence of penalized maximum likelihood problems with self-adaptive penalties that gradually filter out negligible regression coefficients

and partial covariances. Because it adaptively penalizes individual model parameters, our method is seen to outperform fixed-penalty competitors on simulated data. We establish the posterior contraction rate for our model, buttressing our method's excellent empirical performance with strong theoretical guarantees. Using our method, we estimated the direct effects of diet and residence type on the composition of the gut microbiome of elderly adults.

64. A Tangent Approximation approach to Variational Inference in Strongly super-Gaussian likelihood models

[03.M1.I51, (page 27)]

Pritam DEY, *Texas A&M University*
Somjit ROY, *Texas A&M University*
Debdeep PATI, *University of Wisconsin Madison*
Bani MALLICK, *Texas A&M University*

Tangent approximation forms a widely used class of variational inference (VI) methods for Bayesian analysis in intractable, non-conjugate models. These methods leverage convex duality to construct minorants of the marginal likelihood, thereby making inference tractable. However, a general-purpose tangent aprroximation methodology with provable optimality guarantees that extends beyond the commonly studied logit models remains elusive. In this talk, I introduce TAVIE (Tangent Approximation-based Variational InferencE), a general-purpose VI framework tailored for strongly super-Gaussian (SSG) likelihoods, which includes a broad class of non-conjugate models. TAVIE constructs a quadratic lower bound on the log-likelihood, inducing conjugacy with Gaussian priors and enabling scalable variational updates. Critically, we provide theoretical guarantees for the trustworthiness of TAVIE by deriving a variational risk bound with respect to the α-Rényi divergence under the fractional likelihood setup. This result holds under mild conditions on the data-generating process and offers a principled criterion for assessing inference quality. The theoretical insights are further supported by simulation studies and a real data application, demonstrating robust performance across a range of settings.

65 . Propagation of Shocks on Networks: Can Local Information Predict Survival?

[01.A1.I8, (page 10)]

Souvik DHARA, *Purdue University*
Manish PANDEY,
Leonard SCHULMAN, *CalTech*

Complex systems are often fragile, where minor disruptions can cascade into dramatic collapses. Epidemics serve as a prime example of this phenomenon, while the 2008 financial crisis highlights how a domino effect, originating from the small subprime mortgage sector, can trigger global repercussions. The mathematical theory underlying these phenomena is both elegant and foundational, profoundly shaping the field of Network Science since its inception. In this talk, I will present a unifying mathematical model for network fragility and cascading dynamics, and explore its deep connections to the theory of local-weak convergence, pioneered by Benjamini-Schramm and Aldous-Steele.

66. A Branching Process Model for Digital Read Quantification with Application to PCR-Based Diagnostics

[02.M2.I28, (page 19)]

Karin S DORMAN, *Iowa State University*
Debosmita KUNDU, *Iowa State University*

Sequenced read counts are a ubiquitous data summary of modern high throughput biological methods used to observe metagenomes, genomes, transcriptomes, epigenomes, and various kinds of molecular interactions and functions. Almost all such count data are obtained after amplification of sampled molecules, which can bias and overdisperse the biological signal of interest. We develop and investigate a novel model for count data that better adheres to the experimental generative process than Poisson and Negative Binomial models with or without zero-inflation. Our model is based on a Branching Process model of Polymerase Chain Reaction (PCR) amplification. It naturally accounts for overdispersion and zero inflation, with meaningful parameters directly linked to biological processes. In particular, we provide the first estimates of PCR amplification efficiency during library preparation and estimate the effects of primer mismatch on sampling efficiency.

67. Geometric Exploration of Random Objects using Distance Profiles

[01.M2.I3, (page 4)]

Paromita DUBEY, *University of Southern California*
Yaqing CHEN, *Rutgers University*
Hans-Georg MÜLLER, *UC Davis*

In this talk I will propose new tools for the exploratory data analysis of data objects taking values in a general separable metric space. Using distance profiles, where the distance profile of a point ω in the metric space refers to the distribution of the distances between ω and the data objects, I will describe how

to obtain transport ranks, which capture the centrality of each element in the metric space with respect to the data cloud. I will discuss the properties of transport ranks and show how they can be an effective device for detecting and visualizing patterns in samples of random objects. Together with practical illustrations I will establish the large sample properties of the estimators of the distance profiles and the transport ranks which will be valid for a wide class of metric spaces. Finally, I will describe a new powerful two sample test geared towards populations of random objects by utilizing the distance profiles corresponding to the data objects. I will demonstrate the efficacy of this new approach on distributional data comprising of a sample of age-at-death distributions for various countries, for compositional data through energy usage for the U.S. states and for neuroimaging network data.

68 . High-dimensional Asymptotics of Differentially Private PCA

[01.E1.I17, (page 13)]

Rishabh DUDEJA, *UW Madison*
Youngjoo YUN, *UW Madison*

We study the problem of constructing differentially private approximations to the principal components (PCs) of a high-dimensional dataset consisting of n samples, each represented by a p-dimensional feature vector. The exponential mechanism constructs private approximations for the PCs by sampling the privatized PCs from a natural Gibbs distribution defined by the sample covariance matrix. This introduces random noise to the PCs, making it difficult for an adversary to detect the presence or absence of a target individual in the dataset. However, existing privacy analyses of the exponential mechanism are pessimistic and introduce large amounts of noise to meet the desired privacy level, often overwhelming the meaningful signal in the PCs. We present a new analysis of the exponential mechanism that precisely characterizes the information leaked (privacy loss) by the privatized PCs at a given noise level in the high-dimensional asymptotic regime $p \to \infty$. Our result is asymptotically sharp and characterizes the minimum noise required to achieve a target privacy level, resulting in significantly less noisy privatized PCs. We obtain our results via an algorithmic approach that exploits a simple sampling algorithm for the Gibbs distribution to analyze the asymptotic properties of the exponential mechanism.

69. Matrix-free Conditional Simulation of Gaussian random fields

[02.E1.I45, (page 24)]

Somak DUTTA, *Iowa State University*
Debashis MONDAL, *Washington University, St. Louis, MO, USA*

In spatial analysis, conditional simulation of spatial variables at unobserved locations given the data at the observed location facilitates various statistical inferences but suffers from computational scalability when the sample size is large. In this paper, we develop a method for conditional simulation based on novel mathematical decompositions of the inverse-covariance matrix. The method applies to a broad class of spatial models, including the Gaussian Markov random fields, fractional Gaussian fields, and the Matérn models. Matrix-free computational techniques are also developed for scalability. I will describe a practical application to mapping groundwater arsenic exceedance regions.

70 . Are ML methods for supervised classification truly better learners than classical linear discriminant analysis?

[Student Poster Competition, (page 7)]

Oluwafunmibi FASANYA, *University of Nebraska Lincoln*

The Increasing availability of large datasets with limited documentation or anonymized participant/observational unit poses a serious threat since such data could be repeatedly collected from an individual but since detailed documentation or user information is hidden due to privacy, data analyst may mistakenly treat these repeated observations as independent observations. Analyzing such data as an independent measurements without accounting for its longitudinal structure can introduce serious bias in the result in some situations. In this study, we conducted a simulation based study to understand the pitfall of modelling repeated measure data as an independent observations using parametric (LDA) and non-parametric machine learning (ML) algorithms: Support Vector Machine (SVM), Random Forest (RF), Neural Network, and K-Nearest Neighbor (KNN). We evaluated the amount of bias that could result when assuming independence for longitudinal datasets when using LDA and ML models with varying longitudinal correlation structure (CS and AR1), with serial correlations (0, 0.2, 0.5, 0.9, 0.95, and 0.99), and between variable correlation (0.2, 0.5, and 0.9) with different variance for each of the variables. The result showed that for low correlation between the variables, high variances and especially high serial correlation, ML models accuracy shoots high, not because they are learning the struc-

ture of the data better as their value went way over the true accuracy but because they are over predicting the data due to the false pattern introduced by correlation among the observations and structure of the data. However LDA remained stable in its result irrespective of the serial correlation level among the observation, indicating that higher accuracy does not necessarily implies better learning or generalization. These result highlights the importance of paying close attention to how the data was generated and also correctly modeling the data structure to avoid overconfident conclusion. It also stresses that correctly modeling the data structure is not optional but very essential for valid inference and trustworthy predictions.

71 . Optimization-based Sensitivity Analysis

[01.A1.I6, (page 9)]

Tobias FREIDLING, *École polytechnique fédérale de Lausanne*
Qingyuan ZHAO, *University of Cambridge*

Causal inference necessarily relies upon untestable assumptions; hence, it is crucial to assess the robustness of obtained results to violations of identification assumptions. However, such sensitivity analysis is only occasionally undertaken in practice, as many existing methods only apply to relatively simple models and their results are often difficult to interpret. We take a more flexible approach to sensitivity analysis and view it as a constrained stochastic optimization problem.

In this work, we address both theoretical and practical challenges. We describe a general framework to conduct sensitivity analysis, study the conditions that allow for consistent estimation in the resulting partially identified model and suggest a bootstrap approach to construct sensitivity intervals.

Then, we employ our proposed procedure to conduct sensitivity analysis for a linear causal effect when an unmeasured confounder and a potential instrument are present. In this common setting, the bias of the OLS and TSLS estimands can be expressed in terms of partial correlations and we show how practitioners can specify sensitivity models that are intuitive to them. Furthermore, we provide several user-friendly visualization tools to assess and interpret the results of the sensitivity analysis and demonstrate our methods on a real study in labour economics.

72. (Bayesian) meta-analysis: statistical methods and their applications in clinical medicine

[Special Invited Session 3, (page 20)]

Tim FRIEDE, *University Medical Center Göttingen*

Meta-analyses of clinical trials are a cornerstone of evidence-based medicine. In clinical medicine, often only a small number of studies, say $2-5$, are available to address a specific research question. In these settings substantial uncertainty is attached to estimates of between-trial heterogeneity in treatment effects. However, standard methods fail to account for this uncertainty resulting in coverage probabilities well below the nominal level for confidence intervals of the overall treatment effect (Bender et al, 2018). We start by reviewing frequentist approaches that account appropriately for this uncertainty by rescaling the standard error and use of t-quantiles rather than normal quantiles in the construction of the confidence intervals (Hartung und Knapp, 2001a,b; Röver et al, 2015). As an alternative we consider Bayesian approaches to random-effects meta-analysis (Friede et al, 2017) and consider practical aspects of their implementation including the choice of priors (Röver et al, 2021, 2023) and the role of trace plots in their interpretation (Röver et al, 2024). Finally, we discuss how predictive distributions and shrinkage estimators can be used to facilitate the integration of data from different sources such as a randomized controlled trial (RCT) and real world data (RWD) such as clinical registries (Röver & Friede, 2020; Röver & Friede, 2025).

73. Bayesian Mixture Models, Non-local Prior Formulations and MCMC Algorithms

[03.M1.I48, (page 27)]

Jairo Alberto FUQUENE PATINO, *Department of Statistics, UC Davis*
Mark STEEL, *University of Warwick*
David ROSSELL, *Universitat Pompeu Fabra in Barcelona*

In this talk I will present the use of Bayesian mixture models in practical settings. I will also discuss Non-local prior alternatives for mixture distributions and Markov Chain Monte Carlo algorithms for posterior inference and model selection. This talk is motivated with real examples.

74. Optimizing Combination Therapies Using a Bayesian Adaptive Design with a Two-dimensional NDLM

[03.M1.I50, (page 27)]

Byron GAJEWSKI, *University of Kansas Medical Center*

The 2023 American Heart Association Guidelines have identified combination therapies as an important knowledge gap and an area of future research likely to offer the best chance of success for delayed cerebral ischemia (DCI) in patients with an aneurysmal spontaneous subarachnoid hemorrhage (aSAH). In this talk, we present an optimized Bayesian adaptive design to identify the best combination of Cilostazol and Human Albumin using a two-dimensional normal dynamic linear model. This design is shown to be smaller, stronger, faster, and benefit more trial participants than fixed and adaptive designs that use an independent model. Further, the two-dimensional approach avoids the difficulty of prespecifying the order of combination therapies required in a one-dimensional normal dynamic linear model.

75 . Statistical Learning of SDEs: A Journey with Riten Mitra

[02.M1.I20, (page 15)]

Arnab GANGULY, *Associate Professor, Department of Mathematics, Louisiana State University*

TBA

76 . Bayesian nonparametric common atoms approach for creating synthetic controls in early-phase glioblastoma trials

[Student Poster Competition, (page 7)]

Bhanu GARG, *University of Texas at Dallas*
Noirrit Kiran CHANDRA,
Lorenzo TRIPPA,
Peter MÜLLER,
Rifaquat Musaffa RAHMAN,
John DE GROOT,

We introduce a model-based system to complement treatment-only glioblastoma (GBM) trials with a synthetic control arm based on external data, using control arms from earlier studies. The main innovations of the proposed system are a model-based approach that accounts for all relevant uncertainties, implied inference on homogeneous patient subpopulations and the use of possibly varying subset of covariates. As a case study, we considered a hypothetical repeat of the InSIGht study by augmenting the control arm within the study with a synthetic control cohort. We also developed a R shiny app to implement the propose approach, using inference that avoids sharing actual patient records.

77. "Studying algorithmic errors in diagnostic and predictive models in AI-Driven Healthcare Systems: A focus on error detection, mitigation and synthetic data generation"

[03.A1.C2, (page 33)]

Isaac GBENE, *South Dakota State University*

The integration of artificial intelligence (AI) in healthcare systems offers transformative potential, enhancing diagnostic accuracy, predictive capabilities, and patient care delivery. However, these advancements are hindered by various forms of errors, including training data errors, problems in sampling, design, labeling, feature, temporal, human-AI-interaction bias, and evaluation metrics used. This study examines the sources and impacts of these problems, with a focus on their implication for health outcome research and ethical considerations. Through an extensive literature review, the study reveals that biases in AI systems often arise from unrepresentative datasets used in training models, flawed model designs, and evolving societal trends, leading to disparities in healthcare outcomes. Mitigation strategies such as improving data representativeness, ensembled methods, and incorporating continuous error detection frameworks are analysed. In this study, several state-of-the-art synthetic data generation methods are investigated and used to generate data to study errors due to evaluation metrics and mitigation methods.

78. Fast and Accurate Fourier Analysis from Irregularly Sampled Data

[02.E1.I45, (page 25)]

Christopher GEOGA, *University of Wisconsin-Madison*
Paul BECKMAN,

In this work, I will introduce and analyze a new method for performing spectral analysis on fully irregularly sampled data. This tool offers several significant improvements over current popular alternatives: it significantly reduces bias, which almost immediately dominate the signal in cases of severely irregular sampling for several existing tools, and it sidesteps numerical issues that come from poor conditioning of the nonuniform Fourier matrix. We further provide concrete theory and guidelines on design parameter choices that provide direct control of the tradeoff between bias and how far into the tails of the spectral density one can look. We will close with demonstrations in one and two dimensions and a discussion of applications and future extensions.

79. PLRD: Partially Linear Regression Discontinuity Inference

[Student Poster Competition, (page 7)]

Aditya GHOSH, *Stanford University*
Aditya GHOSH, *Stanford University*
Guido IMBENS, *Stanford University*
Stefan WAGER, *Stanford University*

Regression discontinuity designs have become one of the most popular research designs in empirical economics. We argue, however, that widely used approaches to building confidence intervals in regression discontinuity designs exhibit suboptimal behavior in practice: In a simulation study calibrated to high-profile applications of regression discontinuity designs, existing methods either have systematic undercoverage or have wider-than-necessary intervals. We propose a new approach, partially linear regression discontinuity inference (PLRD), and find it to address shortcomings of existing methods: Throughout our experiments, confidence intervals built using PLRD are both valid and short. We also provide large-sample guarantees for PLRD under smoothness assumptions.

80. Reproducibility in Statistics

[Plenary Lecture 1, (page 3)]

Debashis GHOSH, *University of Colorado Anschutz Medical Campus*

With the advent of big data and large-scale computational technologies, increasing focus has been paid to the consideration and evaluation of reproducibility in statistics, machine learning and data mining. We review the issues surrounding reproducibility and describe a framework to modeling reproducibility that leverages ideas from multiple testing. Both applied and more theoretical aspects are provided. Time permitting, we will describe a conceptual approach to reproducibility considerations with large-language models.

81. Polyspectral Mean Estimation of General Nonlinear Processes

[01.A1.I7, (page 10)]

Dhrubajyoti GHOSH, *Duke University*
Tucker MCELROY, *U.S. Census Bureau*
Soumendra LAHIRI, *Washington University in St. Louis*

Higher-order spectra (or polyspectra), defined as the Fourier Transform of a stationary process' autocumulants, are useful in the analysis of nonlinear and non Gaussian processes. Polyspectral means are weighted averages over Fourier frequencies of the polyspectra, and estimators can be constructed from analogous weighted averages of the higher-order periodogram (a statistic computed from the data sample's discrete Fourier Transform). We derive the asymptotic distribution of a class of polyspectral mean estimators, obtaining an exact expression for the limit distribution that depends on both the given weighting function as well as on higher-order spectra. Secondly, we use bispectral means to define a new test of the linear process hypothesis. Simulations document the finite sample properties of the asymptotic results. Two applications illustrate our results' utility: we test the linear process hypothesis for a Sunspot time series, and for the Gross Domestic Product we conduct a clustering exercise based on bispectral means with different weight functions.

82. Online Bayesian Variable Selection for Logistic Regression Models With Streaming Data

[02.E1.I44, (page 24)]

Joyee GHOSH, *The University of Iowa*
Shamriddha DE, *The University of Iowa*
Payel GHOSAL, *University of Wisconsin–Madison*

In several modern applications data are generated continuously over time, such as data generated from smartwatches. We assume data are collected and analyzed sequentially in batches. We develop an online Bayesian model selection method for logistic regression, where the selected model can potentially change throughout the data collection process. We use simulation studies to show that our new method can outperform some existing methods. We apply our method to a traffic crash dataset which has been previously analyzed in the context of streaming data.

83. Optimal Estimation and Testing under Horseshoeplus Priors

[03.M1.I49, (page 27)]

Malay GHOSH, *University of Florida*
Zikun QIN,

The paper considers exact optimality of the horseshoe+ priors in estimation and multiple testing of multivariate normal means under sparsity. First, the posterior means under the horseshoe+ prior are shown to be minimax as point estimates of the multivariate normal means under sparsity. Then, under a thresholding multiple testing procedure, the horseshoe+ prior is shown to attain asymptotic Bayes optimality again under sparsity. We further propose an empirical Bayes approach for the multiple testing problem and demonstrate its optimality. The paper

also includes some discussions on its connection to the existing literature.

84. Signal-to-noise ratio aware minimax analysis of sparse linear regression

[Student Paper Competition 1, (page 5)]

Shubhangi GHOSH, *Columbia University*
Yilin GUO,
Haolei WENG,
Arian MALEKI, *Columbia University*

We consider parameter estimation under sparse linear regression – an extensively studied problem in high-dimensional statistics and compressed sensing. While the minimax framework has been one of the most fundamental approaches for studying statistical optimality in this problem, we identify two important issues that the existing minimax analyses face: (i) The signal-to-noise ratio appears to have no effect on the minimax optimality, while it shows a major impact in numerical simulations. (ii) Estimators such as best subset selection and Lasso are shown to be minimax optimal, yet they exhibit significantly different performances in simulations. In this paper, we tackle the two issues by employing a minimax framework that accounts for variations in the signal-to-noise ratio (SNR), termed the SNR-aware minimax framework. We adopt a delicate higher-order asymptotic analysis technique to obtain the SNR-aware minimax risk. Our theoretical findings determine three distinct SNR regimes: low-SNR, medium-SNR, and high-SNR, wherein minimax optimal estimators exhibit markedly different behaviors. The new theory not only offers much better elaborations for empirical results, but also brings new insights to the estimation of sparse signals in noisy data.

85. A Robust Kernel Machine Framework for Assessing Spatial Variability and Cross-Niche Communication in Spatial Transcriptomics

[02.M2.I27, (page 18)]

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

In this talk, we present CytoKSpace, a robust kernel-based approach that simultaneously identifies spatially variable genes (SVGs) and detects cross-niche ligand-receptor communication in spatial transcriptomics. Our method accommodates large-scale single-cell or spot-level data by leveraging low-rank approximations of kernel embeddings, enabling efficient inference on tissue-wide expression patterns. It further incorporates niche-level adjacency to reveal potential ligand-receptor interactions among adjacent tissue compartments. We introduce two permutation schemes—shuffling expression labels or cluster assignments- to rigorously assess null hypotheses around spatial dependence and cluster relevance. Through extensive benchmarks on synthetic and real spatial transcriptomics datasets, CytoKSpace effectively controls the false discovery rate (FDR) while identifying spatially variable patterns and biologically meaningful cross-niche signals. We demonstrate its capacity to pinpoint functionally important ligand-receptor pairs across neighboring niches, advancing critical insights into tissue microenvironments and cell-cell communication.

86 . Oracle optimal unsupervised Bayesian image segmentation

[03.M1.I47, (page 26)]

Subhashis GHOSHAL, *North Carolina State University*
Eduard BELITSER, *VU Amsterdam*
Shuvrarghya GHOSH, *North Carolina State University*

Identifying homogeneous parts of an image as certain objects is an important goal in image analysis and computer vision applications. Most image segmentation methods are designed to work under a supervised learning framework and need a vast amount of data to train. We consider a simple unsupervised image segmentation method based on a tree partitioning structure and construct a prior distribution in a Bayesian framework. We obtain local (oracle) rates for estimation and prediction and show that the Bayesian procedure mimics the oracle rate. Further, we show that the Bayesian procedure recovers the segments in terms of the Rand index for clustering under a mild condition on the differences of signal values at different segments. We discuss a Markov chain-based computational method to obtain image segments. We illustrate the method on some test images.

87. HEART: Heterogeneous data-driven Emotion and Anomaly Recognition in Sparse Longitudinal Texts

[02.M1.I24, (page 17)]

Aritra GUHA, *AT&T Chief Data Office*
Prasanjit DUBEY, *Georgia Institute of Technology*
Paromita DUBEY, *University of Southern California*
Zhengyi ZHOU, *AT&T*

Sparse longitudinal data arise in diverse applications, such as market research, medical studies, and social sciences. These data are complex due

to sparse and irregularly spaced observations over time for each subject. Functional data analysis provides a nonparametric framework for analyzing such data; however, existing methods do not accommodate non-tabular data formats, such as textual information. Downstream tasks like anomaly detection pose significant challenges in the context of sparse longitudinal data, particularly when measurement errors and subject-level heterogeneity are present. These challenges become even more pronounced when the subject-specific measurements involve textual data. For example, in market research, detecting outlying patterns in customer emotion journeys is a critical task, often relying on limited customer feedback or interactions within a given time period. This work introduces a novel framework that combines sparse Functional Principal Component Analysis (sFPCA) with covariate-informed iterative clustering to analyze sparsely observed, longitudinal, and heterogeneous text data. Leveraging GPT-3.5 Turbo, we detect emotions based on Plutchik's wheel and develop statistical tests with theoretical guarantees to identify anomalous trajectories. We demonstrate the utility of our method on Amazon customer reviews, applying the framework to pinpoint critical pain points in the customer journey. Our approach uncovers anomalous patterns in customer experiences, providing actionable insights for data-driven decision-making in consumer analytics.

88. Bayes in Multi-Layer Networks

[03.E1.I67, (page 35)]

Sharmistha GUHA, *Texas A&M University*

We present an approach for analyzing multilayer networks to address complex inference challenges in fields like security and neuroscience. We introduce a supervised learning framework that leverages inter- and intra-layer dependencies to predict continuous outcomes. Using low-rank models, this method captures intricate relationships, identifies key nodes and edges, and improves computation speed. Its effectiveness is demonstrated on network security data from a national laboratory, significantly enhancing prediction accuracy.

89. Bayesian Estimation of Propensity Scores for Integrating Multiple Cohorts with High-Dimensional Covariates

[02.A1.I36, (page 22)]

Subharup GUHA, *University of Florida*
Subharup GUHA, *University of Florida*
Yi LI, *University of Michigan*

Comparative meta-analyses of groups of subjects by integrating multiple observational studies rely on estimated propensity scores (PSs) to mitigate covariate imbalances. However, PS estimation grapples with the theoretical and practical challenges posed by high-dimensional covariates. Motivated by an integrative analysis of breast cancer patients across seven medical centers, this paper tackles the challenges of integrating multiple observational datasets. The proposed inferential technique, called Bayesian Motif Submatrices for Covariates (B-MSC), addresses the curse of dimensionality by a hybrid of Bayesian and frequentist approaches. B-MSC uses nonparametric Bayesian "Chinese restaurant" processes to eliminate redundancy in the high-dimensional covariates and discover latent motifs or lower-dimensional structures. With these motifs as potential predictors, standard regression techniques can be utilized to accurately infer the PSs and facilitate covariate-balanced group comparisons. Simulations and meta-analysis of the motivating cancer investigation demonstrate the efficacy of the B-MSC approach to accurately estimate the propensity scores and efficiently address covariate imbalance when integrating observational health studies with high-dimensional covariates and right-censored survival outcomes.

90. Totally concave regression

[Special Invited Session 2, (page 17)]

Adityanand GUNTUBOYINA, *University of California Berkeley*
Dohyeong KI, *University of California Berkeley*

Shape constraints in nonparametric regression provide a powerful framework for estimating regression functions under realistic assumptions without tuning parameters. However, most existing methods–except additive models–impose too weak restrictions, often leading to overfitting in high dimensions. Conversely, additive models can be too rigid, failing to capture covariate interactions. This paper introduces a novel multivariate shape-constrained regression approach based on total concavity, originally studied by T. Popoviciu. Our method allows interactions while mitigating the curse of dimensionality, with convergence rates that depend only logarithmically on the number of covariates. We characterize and compute the least squares estimator over totally concave functions, derive theoretical guarantees, and demonstrate its practical effectiveness through empirical studies on real-world datasets. This is joint work with Dohyeong Ki from UC Berkeley.

91. Bayesian Ordinal Network Meta-Regression under General Links with Applications to Crohn's Disease

[04.M2.I73, (page 37)]

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

Logistic regression models are widely used for modeling ordinal response data due to their attractive proportional odds property and computational efficiency. However, adequacy of such model has not been well examined. In this paper, we present a meta-regression approach for modeling aggregate ordinal outcomes using different links, beyond the logit link. Specifically, we use a regression model based on aggregate covariates to model cut points and the variance of trial-level random effects to account for heterogeneity across studies. We also investigate an importance of the links, which leads to improved model fitting and evaluation. Our theoretical finding allows for the incorporation of a variety of links, regardless of symmetry or asymmetry. Additionally, we develop an efficient Markov chain Monte Carlo sampling algorithm to perform Bayesian computation under different links. We further propose a new concordance index to evaluate the performance of the model in fitting aggregate ordinal outcomes. Two Bayesian model comparison measures, the DIC and WAIC, are used to determine the appropriateness of links and assess goodness-of-fit. We carry out an extensive simulation study to examine the empirical performance of the proposed model under different links. A case study is presented to demonstrate the usefulness of the proposed methodology, using aggregate ordinal outcome data from 16 clinical trials for treating Crohn's disease.

92. When few labeled target data suffices: a theory of semi-supervised domain adaptation via fine-tuning from multiple starts

[02.E1.I43, (page 24)]

Wooseok HA, *KAIST*
Yuansi CHEN, *ETH Zurich*

Semi-supervised domain adaptation (SSDA) aims to achieve high predictive performance in the target domain with limited labeled target data by making use of abundant source and unlabeled target data. Despite its practical significance in numerous applications, the theoretical understanding of SSDA methods remains largely unexplored, particularly in scenarios involving different types of distributional shifts between source and target data. In this work, we develop a theoretical framework based on structural causal models to analyze and quantify the performance of SSDA algorithms when labeled target data is limited. We introduce fine-tuning strategies tailored to distinct assumptions about the relationship between source and target distributions and show how these methods extend unsupervised domain adaptation (UDA) models trained on source and unlabeled target data to achieve higher target performance with lower target sample complexity. When the relationship between source and target data is vaguely known—a common scenario in real-world applications—we propose the Multi-Start Fine-Tuning (MSFT) algorithm, which fine-tunes UDA models using multiple starting points and selects the best-performing model based on a small hold-out target validation data. By combining theoretical guarantees for individual fine-tuning strategies with model selection, MSFT achieves near-optimal target predictive performance across a broad range of types of distribution shifts while significantly reducing the need for labeled target samples compared to methods on target data alone. We demonstrate the effectiveness of our proposed algorithms through simulations to complement our theoretical analysis.

93. Fiducial Generative Models

[Special Invited Session 5, (page 30)]

Jan HANNIG, *University of North Carolina at Chapel Hill*
Zijie TIAN, *UC Davis*
Thomas C.M. LEE, *tcmlee@ucdavis.edu*

While generalized fiducial inference (GFI) and its variants have yielded many theoretical and practical results to parametric inference and uncertainty quantification, applying it to generative models remains challenging. We identify three key issues misspecification, metric choices, and over-parameterization hinder the direct application of the GFI to generative models. In this paper, we propose a novel method based on the framework of generalized fiducial inference, designed to construct distributional estimates over the parameter space given observed data, while also enabling uncertainty quantification for generative models. We employ a truncation-based approach and further provide a theoretical analysis of its behavior under varying truncation parameters. Both theoretical results and empirical evidence suggest that, with an appropriately chosen truncation parameter, the truncated distribution derived from

generalized fiducial inference achieves valid coverage of the true parameter and leads to improved generalization performance.

94. Evaluation of Cox Mixture Models for End-Stage Kidney Disease for Higher Risk Patients

[Student Poster Competition, (page 7)]

Jason R HASSE, *South Dakota State University*
Semhar MICHAEL,

Persons with end-stage kidney disease (ESKD) have a substantially impacted quality of life requiring frequent dialysis or a kidney transplant. Due to the large range of socio-economic factors and demographics in the United States (US), the assumption of proportional hazards (PH) which is required for Cox Regression, could be violated. To remedy this violation, an investigation into the appropriate subpopulations which better satisfy the PH assumption is performed. Data from the United State Renal Data System (USRDS) on patients with ESKD is analyzed. Cox mixtures (CM) and deep Cox mixtures (DCM) models are utilized to identify and model the latent subpopulations while simultaneously modeling time to death. CM models maintain the interpretability of the typical Cox regression model with the increased performance of the mixture model. We found that both CM and DCM models outperform the Cox model in terms of Brier score and a time-dependent concordance index. The analysis also showed varied performance by race/ethnicity and geographic subpopulations to study disparities.

95. What does Guidance do in Masked Discrete Diffusion Models

[01.A1.I7, (page 10)]

Ye HE, *Georgia Institute of Technology*
Kevin ROJAS, *Georgia Institute of Technology*
Molei TAO, *Georgia Institute of Technology*

We study masked discrete diffusion models with classifier-free guidance (CFG) under an absorbing forward process. Assuming no score or discretization error, we derive an explicit solution for the guided reverse dynamics and analyze how guidance influences the sampling behavior. When the full data distribution is a mixture over classes and the goal is to sample from a specific class, guidance amplifies class-specific regions while suppresses regions shared with other classes. This effect depends on the guidance parameter w and induces distinct covariance structures in the sampled distribution. Notably, we observe quantitatively different behaviors in 1D and 2D. We also show that the convergence rate of the reverse dynamics also depends on w, with different scaling behaviors in 1D and 2D. These findings highlight the role of guidance not just in shaping the output distribution but also in controlling the dynamics of the sampling trajectory. Our theoretical analysis is supported by experiments that illustrate the geometric effects of guidance and its impact on convergence.

96. Efficient Analysis of Latent Spaces in Heterogeneous Networks

[02.E1.I38, (page 22)]

Yinqiu HE, *University of Wisconsin-Madison*
Tian YUANG, *Fudan University*
Jiajin SUN, *Florida State University*

In this talk, we will discuss a unified framework for efficient estimation under latent space modeling of heterogeneous networks. We consider a class of latent space models that decompose latent vectors into shared and network-specific components across networks. We develop a novel procedure that first identifies the shared latent vectors and further refines estimates through efficient score equations to achieve statistical efficiency. Oracle error rates for estimating the shared and heterogeneous latent vectors are established simultaneously. The analysis framework offers remarkable flexibility, accommodating various types of edge weights under exponential family distributions.

97. Enhancing High-Dimensional Time Series Analysis with Envelope Methods

[03.A1.I60, (page 32)]

Wiranthe HERATH, *Drake University*
Yaser SAMADI, *Southern Illinois University Carbondale*

Standard vector autoregressive (VAR) models face significant challenges in high-dimensional settings due to overparameterization, which limits the incorporation of multiple variables and lags. Existing dimension reduction approaches—such as reduced-rank and envelope-based methods—provide valuable tools for addressing these challenges, yet they may be limited in fully capturing the most informative subspace or in effectively managing rank deficiency.

This talk explores a unified framework that leverages the strengths of both methodologies to achieve greater parsimony, improved estimation efficiency, and enhanced forecasting performance. Emphasis will be placed on the theoretical properties of the proposed approach and its empirical advantages in analyzing complex multivariate time series data.

98. Foundation of Mixture of Experts in Large-Scale Machine Learning Models

[03.A1.I59, (page 31)]

Nhat HO, *The University of Texas, Austin*

Mixtures of experts (MoEs), a class of statistical machine learning models that combine multiple models, known as experts, to form more complex and accurate models, have been combined into deep learning architectures to improve the ability of these architectures and AI models to capture the heterogeneity of the data and to scale up these architectures without increasing the computational cost. In mixtures of experts, each expert specializes in a different aspect of the data, which is then combined with a gating function to produce the final output. Therefore, parameter and expert estimates play a crucial role by enabling statisticians and data scientists to articulate and make sense of the diverse patterns present in the data. However, the statistical behaviors of parameters and experts in a mixture of experts have remained unsolved, which is due to the complex interaction between gating function and expert parameters.

In the first part of the talk, we investigate the performance of the least squares estimators (LSE) under a deterministic MoEs model where the data are sampled according to a regression model, a setting that has remained largely unexplored. We establish a condition called strong identifiability to characterize the convergence behavior of various types of expert functions. We demonstrate that the rates for estimating strongly identifiable experts, namely the widely used feed-forward networks with activation functions sigmoid($\cdot$) and tanh($\cdot$), are substantially faster than those of polynomial experts, which we show to exhibit a surprising slow estimation rate.

In the second part of the talk, we show that the insights from theories shed light into improving important practical applications, including enhancing the performance of Transformer model with a novel self-attention mechanism, efficiently finetuning large-scale AI models for downstream tasks, and effectively scaling up massive AI models with several billion parameters.

99. Making prediction intervals smarter: randomization enables local coverage

[02.M1.C1, (page 16)]

Rohan HORE, *University of Chicago*
Rina BARBER, *University of Chicago*

How can we build prediction intervals that are not only valid on average, but also reliable for individuals? Conformal prediction (CP) provides powerful tools for constructing distribution-free prediction intervals with finite-sample guarantees. However, these guarantees are marginal, as they average over all inputs and can leave room for error in specific subpopulations or regions of the data.

In this talk, I will explore how to bridge this gap using randomization and local weighting. We will revisit the challenge of local coverage, and I will introduce Randomly Localized Conformal Prediction (RLCP)—a new method that blends ideas from weighted and localized conformal prediction. RLCP offers relaxed but theoretically grounded local guarantees, and even validity under smooth covariate shifts. Along the way, I will discuss why exact local guarantees are impossible, how randomization helps us get close, and how this leads to improvements in the reliability and fairness of predictive models.

100. Allowing Negative Variance Component Estimates in REML: Inferential Consequences for Fixed Effects and Type I Error Control

[04.M1.I69, (page 36)]

Reka HOWARD, *University of Nebraska-Lincoln*
Bipin POUDEL, *University of Nebraska-Lincoln*
Nora BELLO, *USDA*
Walt STROUP, *University of Nebraska-Lincoln*

This study examines the inferential implications of allowing for negative variance component estimates in mixed model analyses using Restricted Maximum Likelihood (REML). Specifically, we make comparisons with the common practice of constraining variance estimates to be non-negative, thus effectively dropping random effects from the linear predictor when variance estimates are set to zero when estimating equations produce a negative solution. We show this common practice to be ultimately undesirable as it yields a model specification that misrepresents the data generation process. We further evaluate the impact of these constraints on the variance estimates on practical inference in the context of common multilevel experimental designs. Specifically, we explore inferential implications on significance levels for fixed effects and on estimation of standard errors and confidence intervals. Consistent with the limited prior research on this topic, our findings indicate that constraining variance component estimates to be non-negative inflates the Type I error rate when testing the significance of fixed effects. Similar inferential implications are apparent from standard errors and confidence intervals. This study underscores the need for proper model specification that adequately reflects

the data generation process, and its corresponding implementation in statistical software packages. Ultimately, we advocate for best practices in the implementation of mixed model analyses to ensure accurate and reliable inference.

101. Universality Phenomenon in Random Feature and Kernel-based Learning

[01.E1.I17, (page 13)]

Hong HU, *Washington University in St. Louis*

Universality corresponds to the high-dimensional phenomenon that the macroscopic properties of a large-scale system do not depend too much on its microscopic structure. Such phenomenon has been explored and exploited in many different fields such as statistical physics, random matrix theory and signal processing. In this talk, I will present recent works on understanding the exact asymptotics of random feature and kernel-based learning. One major challenge in the analysis stems from the non-linearity of underlying machine learning models. It turns out that in this context, there is a universality phenomenon called Gaussian equivalence: in terms of macroscopic performance such as training or generalization errors, the non-linear models can be equivalent to some microscopically different Gaussian models, which are much easier to analyze. This universality phenomenon enables us to obtain sharp characterization which reveals how the scalings of model and sample sizes, regularization, activation and target functions jointly affect the learning performance.

102. Nonparametric inference on non-negative dissimilarity measures at the boundary of the parameter space

[02.M2.I26, (page 18)]

Aaron HUDSON, *Fred Hutchinson Cancer Center*

It is often of interest to assess whether a function-valued statistical parameter, such as a density function or a mean regression function, is equal to any function in a class of candidate null parameters. This can be framed as a statistical inference problem where the target estimand is a scalar measure of dissimilarity between the true function-valued parameter and the closest function among all candidate null values. These estimands are typically defined to be zero when the null holds and positive otherwise. While there is well-established theory and methodology for performing efficient inference when one assumes a parametric model for the function-valued parameter, methods for inference in the nonparametric setting are limited. When the null holds, and the target estimand resides at the boundary of the parameter space, existing nonparametric estimators either achieve a non-standard limiting distribution or a sub-optimal convergence rate, making inference challenging. In this work, we propose a strategy for constructing nonparametric estimators with improved asymptotic performance. Notably, our estimators converge at the parametric rate at the boundary of the parameter space and also achieve a tractable null limiting distribution. To illustrate, we discuss how this framework can be applied to perform inference in nonparametric regression problems and to perform nonparametric assessment of stochastic dependence.

103. Bayesian structured variable selection in finite mixture of regression analysis for cancer data

[01.A1.I10, (page 11)]

Yunju IM, *University of Nebraska Medical Center*

Cancer is well-known for its heterogeneous nature, which has motivated extensive efforts to identify latent subgroups that exhibit distinct relationships with clinical outcomes. With high-dimensional genetic variables, much of the existing work has focused on identifying subgroup-specific variables, screening out noise. However, this perspective often overlooks the possibility that some effects may be shared across subgroups, reflecting underlying biological similarities. In this project, we develop a novel Bayesian approach based on finite mixture models that allows for simultaneous identification of heterogeneous and homogeneous effects, while effectively removing irrelevant variables. Simulation studies demonstrate that the proposed approach outperforms existing methods in both subgroup detection and variable selection. The analysis of TCGA data on lung cancer data is performed, revealing biologically meaningful subgroup structures and distinct sets of important variables, each with either subgroup-specific or shared effects.

104. Generalizing Regret bounds for Thompson sampling through Minimax-Optimal α-Posterior Concentration Analysis

[01.E1.I18, (page 13)]

Prateek JAISWAL, *Purdue University*
Debdeep PATI, *Department of Statistics, University of Wisconsin-Madison*
Anirban BHATTACHARYA, *Department of Statistics, Texas A&M University*
Bani MALLICK, *Department of Statistics, Texas A&M*

University

This talk introduces a novel analysis of Thompson Sampling (TS) using -posteriors, where the likelihood is tempered by a factor $\alpha \in (0, 1)$. We first present a new α-posterior concentration result that achieves minimax-optimal rates for finite-dimensional models under more refined prior-thickness assumption than previous works.

Leveraging this, we analyze α-Thompson Sampling (α-TS) and derive both instance-dependent and instance-independent regret bounds. Our bounds match existing results and remain valid for broad classes of priors and reward models. This work offers a general framework for regret analysis in Bayesian bandits, resolving key limitations in prior TS analyses that require conjugacy or closed-form posteriors.

105 . Some Mixture models for joint analysis of wind speed and wind direction

[03.A1.C2, (page 33)]

DEBARGHYA JANA, *Iowa State University*
Arnab HAZRA, *Assistant Professor of Statistics at the Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, Kanpur, India.*

In this paper, we propose the Isotropic Gaussian Mixture model(IGM) and Anisotropic Gaussian Mixture model(AGM) for the joint modelling of wind speed and wind direction of cyclones within a semi-parametric framework. We applied the Expectation-Maximization(EM) algorithm to find the estimates of the unknown parameters of our proposed models. The motivation behind writing this paper is to offer crucial support in establishing resilient disaster preparedness measures for cyclones in the North Indian Ocean basin. Especially, the coastal areas in this region are infamous for their recurrent and severe devastation, leading to significant damage, loss of life, and economic upheaval annually due to continuous cyclone activities. Joint modelling of wind speed and wind direction of cyclones plays a crucial role in predicting storm surges and is a critical element in forecasting coastal flooding which is indispensable for government policymakers, as it significantly contributes to the improvement of disaster management in coastal regions. There are some parametric and nonparametric models where the joint model mostly figures out the wind energy analysis As there is no significant literature where these crucial findings are encountered in this context, we introduce the above-mentioned mixture models (AGM and IGM model) and compare their performance using some popular metrics of measuring the goodness of fit, against a few established well-known parametric and non-parametric models which are described in a few existing literature.

106. Can Empirical Bayes via Empirical Risk Minimization Balance Computational and Theoretical Benefits?

[02.A1.I34, (page 21)]

Soham JANA, *University of Notre Dame*
Yury POLYANSKIY, *MIT*
Anzo TEH, *MIT*
Yihong WU, *Yale University*

We study the problems of estimating Poisson means via empirical Bayes modeling. The classical Robbin's method is fast to compute, however, it severely suffers from small sample issues. On the other hand, the nonparametric likelihood based estimators provide decent performances in practice, but they are difficult to compute in multidimensional settings. We propose an empirical risk minimization based approach that provides a balanced solution with respect to both the worlds. We will also discuss generality of our strategy to other setups and practical implications.

107. Scalable divergence time estimation via Hamiltonian Monte Carlo sampling

[02.M2.I28, (page 19)]

Xiang JI, *Tulane University*

Advances in genome sequencing technology are generating genetic data at an ever-increasing pace. This burst of data provides opportunities to look at the underlying biological processes that generate evolutionary patterns. However, these opportunities are accompanied by both statistical and computational challenges that require scalability for making inference with large amount of sequences. In this talk, I will discuss our newly developed ratio transformation and its enabled usage of the Hamiltonian Monte Carlo sampling methods to learn the divergence times of fast-evolving pathogens. With previous approaches, these inferences would have been computationally intractable.

108 . Bayesian Sparse Regression for Microbiome-Metabolite Data Integration

[Student Poster Competition, (page 7)]

Kai JIANG, *The University of Texas Health Science Center at Houston*
Satabdi SAHA, *The University of Texas MD Anderson*

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

Numerous studies have shown that microbial metabolites, which represent the products of bacteria in the human gut, play a key role in shaping cancer risk and response to treatment. However, metabolite data typically contain a large proportion of missing values, which may result from either low abundance or technical challenges in data processing. Moreover, given the compositionality of microbiome data, where the observed abundances can only be interpreted on a relative scale, standard variable selection methods are not applicable. In this project, we propose a novel Bayesian method to address challenges in both metabolite and microbiome data. Key features of our proposed model include adopting a Bayesian prior designed to address the compositional characteristics of microbiome data and modeling the two different mechanisms of missing metabolite data. We demonstrate on simulated data that our proposed model can accurately impute the unobserved true metabolite values and correctly select the relevant microbiome predictors. We illustrate our method using real data from a study on the interplay between the microbiome and metabolome in colorectal cancer.

109 . Instance-optimal stochastic optimization: Succeeding when sample average approximation fails

[02.A1.I35, (page 21)]

Liwei JIANG, *Georgia Institute of Technology*

We consider optimization of a smooth and strongly convex population loss function under a stochastic oracle that can inject both additive and multiplicative noise. Our setting covers quadratic optimization and linear regression as special cases. We begin by establishing finite-sample, information-theoretic local minimax lower bounds in this setting, bringing out the fundamental, instance-dependent aspects of the problem that influence the quality of any optimal solution. With these benchmarks in hand, we then show that both sample average approximation as well as robust (or averaged) stochastic approximation suffer from undesired behavior, in that they are either inconsistent or fail to match our lower bounds in practical sample size regimes. In contrast to these methods, we show that a careful form of variance reduction can perform in an instance-optimal manner while using the minimal possible sample size. Our findings are supported by several careful numerical studies.

110. A Bayesian Generalized Bridge Regression Approach to Covariance Estimation in the Presence of Covariates

[Special Invited Session 5, (page 30)]

Galin JONES, *University of Minnesota*
Christina ZHAO, *AbbVie*
Adam ROTHMAN, *University of Minnesota*

A hierarchical Bayesian approach is proposed that permits simultaneous inference for the regression coefficient matrix and the error precision (inverse covariance) matrix in the multivariate linear model. Assuming a natural ordering of the elements in the response, the precision matrix is reparameterized so that it can be estimated using univariate-response linear regression techniques. A novel generalized bridge regression prior that accommodates both sparse and dense settings, and is competitive with alternative methods for univariate-response regression, is proposed and utilized in this framework. Two component-wise Markov chain Monte Carlo algorithms are developed for sampling, including a data augmentation algorithm based on a scale mixture of normals representation. Numerical examples demonstrate that the proposed method is competitive with comparable joint mean-covariance models, particularly in estimation of the precision matrix. The method is also used to estimate the 253 x 253 precision matrices of two classes of spectra extracted from images taken by the Hubble Space Telescope. Some interesting structural patterns in the estimates are discussed.

111. Prediction of Tropical Pacific Rain Rates with Overparameterized Neural Networks

[03.M1.I46, (page 26)]

Mikyoung JUN, *University of Houston*
Hojun YOU,
Jiayi WANG,
Raymond WONG,

The prediction of tropical rain rates from atmospheric profiles poses significant challenges, mainly due to the heavy-tailed distribution exhibited by tropical rainfall. This study introduces overparameterized neural networks not only to forecast tropical rain rates but also to explain their heavy-tailed distribution. The investigation is separately conducted for three rain types (stratiform, deep convective, and shallow convective) observed by the Global Precipitation Measurement satellite radar over the west and east Pacific regions. Atmospheric profiles of humidity, temperature, and zonal and meridional winds from the MERRA-2 reanalysis are considered

as features. Although overparameterized neural networks are well known for their "double descent phenomenon," little has been explored about their applicability to climate data and capability of capturing the tail behavior of data. In our results, overparameterized neural networks accurately estimate the rain-rate distributions and outperform other machine learning methods. Spatial maps show that overparameterized neural networks also successfully describe the spatial patterns of each rain type across the tropical Pacific. In addition, we assess the feature importance for each overparameterized neural network to provide insight into the key factors driving the predictions, with low-level humidity and temperature variables being the overall most important. These findings highlight the capability of overparameterized neural networks in predicting the distribution of the rain rate and explaining extreme values.

112. Adaptive Divide and Conquer with Two Rounds of Communication

[Student Poster Competition, (page 7)]

Niladri KAL, *Texas A&M University*
Debdeep PATI, *University of Wisconsin-Madison*
Botond SZABO, *Bocconi University*
Rajarshi GUHANIYOGI, *Texas A&M University*

Existing adaptive, distributed computing methods restrict either the number of machines or the smoothness range over which adaptation is possible, limiting their potential applicability. We introduce a novel two-round communication strategy that enables adaptive, rate-optimal estimation without such stringent restrictions. In the first round, local machines send summary statistics to the central node to estimate the hyperparameters that are dependent on the underlying smoothness. In the second round, another set of statistics are transmitted, enabling the central machine to aggregate and produce a final estimator that adapts to the true smoothness level. This approach achieves optimal convergence rates across a wider range of regularities, offering a potential improvement in the adaptability and efficiency of distributed estimation.

113. Power properties of the two-sample test based on the nearest neighbors graph

[Student Paper Competition 1, (page 5)]

Rahul Raphael KANEKAR, *Stanford University*

In this paper, we study the problem of testing the equality of two multivariate distributions. One class of tests used for this purpose utilizes geometric graphs constructed using inter-point distances. So far, the asymptotic theory of these tests applies only to graphs which fall under the stabilizing graphs framework of Penrose and Yukich. We study the case of the K-nearest neighbors graph where $K = k\_N$ increases with the sample size, which does not fall under the stabilizing graphs framework. Our main result gives detection thresholds for this test in parametrized families when $k\_N = o(N^{1/4})$, thus extending the family of graphs where the theoretical behavior is known. We propose a 2-sided version of the test which removes an exponent gap that plagues the 1-sided test. Our result also shows that increasing the number of nearest neighbors boosts the power of the test. This provides theoretical justification for using denser graphs in testing equality of two distributions.

114. Quasi-Bayes in Conditional Moment Restriction Models

[03.M1.I51, (page 27)]

Sid KANKANALA, *University of Chicago*

This paper develops a quasi-Bayes framework for nonparametric structural functions that are identified through a conditional moment restriction. We establish contraction rates for a class of Gaussian process priors and provide conditions under which a Bernstein–von Mises theorem holds for the quasi-posterior distribution. Consequently, we demonstrate that optimally weighted quasi-Bayesian credible sets achieve exact asymptotic frequentist coverage. This extends classical results on the frequentist validity of optimally weighted quasi-Bayesian credible sets in parametric generalized method of moments (GMM) models.

115. Risk-inclusive Contextual Bandits for Early Phase Clinical Trials

[Student Paper Competition 2, (page 5)]

Rohit KANRAR, *Iowa State University*
Chunlin LI,
Zara GHODSI,
Margaret GAMALO,

Early-phase clinical trials face the challenge of selecting optimal drug doses that balance safety and efficacy due to uncertain dose-response relationships and varied participant characteristics. Traditional randomized dose allocation often exposes participants to sub-optimal doses by not considering individual covariates, requiring larger sample sizes and prolonging drug development. This paper introduces a risk-inclusive contextual bandit algorithm that utilizes multi-arm bandit (MAB) strategies, previously suc-

cessful in recommendation systems, to optimize dosing through participant-specific data integration. By combining two separate Thompson samplers—one for efficacy and one for safety, the algorithm enhances the balance between efficacy and safety in dose allocation. The effect sizes are estimated with a generalized version of asymptotic confidence sequences (AsympCS) (Waudby-Smith et al., 2024a), offering a uniform coverage guarantee for sequential causal inference over time. The validity of AsympCS is also established in the MAB setup. The empirical results demonstrate the superiority of this method in optimizing dose allocation compared to randomized allocations and traditional contextual bandits focused solely on efficacy. Moreover, an application on real data generated from a recent Phase IIb study aligns with actual findings.

116. Optimal Sequential Recommendation Systems

[03.E1.I64, (page 34)]

Mina KARZAND, *UC Davis*
Mina KARZAND, *Department of Statistics, UC Davis*
Guy BRESLER, *EECS Department, MIT*

We consider an online model for recommendation systems, with each user being recommended an item at each time-step and providing 'like' or 'dislike' feedback. A latent variable model specifies the user preferences: both users and items are clustered into types. The model captures structure in both the item and user spaces, as used by item-item and user-user collaborative filtering algorithms. We study the situation in which the type preference matrix has i.i.d. entries. Our main contribution is an algorithm that simultaneously uses both item and user structures, proved to be near-optimal via corresponding information-theoretic lower bounds. In particular, our analysis highlights the sub-optimality of using only one of item or user structure (as is done in most collaborative filtering algorithms).

117. Quantifying Uncertainty in Crop Yield Predictions using Deep Learning Ensembles for Risk-Informed Decision-Making

[01.A1.I9, (page 11)]

Venkata Sai Pramod Kumar KASTURI, *Corteva Agriscience*
Bishwa SAPKOTA, *Corteva Agriscience*
Venkat NEMANI, *Corteva Agriscience*
Hoda HELMI, *Corteva Agriscience*

The agriculture industry is significantly impacted by uncertainty arising from environmental factors, genetic variability, and management practices. Accurate prediction of crop yield, particularly corn yield, is crucial for optimizing resource allocation, planning harvests, and making informed decisions. In this presentation, we present a novel approach to estimate corn yield along with uncertainty using a weighted ensemble of diverse uncertainty-based models. These models are optimized to give high quality of uncertainty by balancing out over- and under-confidence through the concept of calibration curve. This enables the stakeholders to make well-informed risk-based decisions. This framework not only enables increased reliability of yield predictions but is also expanded into classification of fields based on productivity.

118 . Inference with Gromov-Wasserstein Distances

[Special Invited Session 2, (page 17)]

Kengo KATO, *Cornell University*
Gabriel RIOUX, *Cornell University*
Ziv GOLDFELD, *Cornell University*

The Gromov-Wasserstein (GW) distance enables comparing metric measure spaces based solely on their internal structure, making it invariant to isomorphic transformations. This property is particularly useful for comparing datasets that naturally admit isomorphic representations, such as unlabelled graphs or objects embedded in space. However, apart from the recently derived empirical convergence rates for the quadratic GW problem, a statistical theory for valid estimation and inference remains largely obscure. Pushing the frontier of statistical GW further, this work derives the first limit laws for the empirical GW distance across several settings of interest: (i) discrete, (ii) semi-discrete, and (iii) general distributions under moment constraints under the entropically regularized GW distance. The derivations rely on a novel stability analysis of the GW functional in the marginal distributions. The limit laws then follow by an adaptation of the functional delta method. As asymptotic normality fails to hold in most cases, we establish the consistency of an efficient estimation procedure for the limiting law in the discrete case, bypassing the need for computationally intensive resampling methods. We apply these findings to testing whether collections of unlabelled graphs are generated from distributions that are isomorphic to each other.

119. Bridging Root-nand Non-standard Asymptotics: Adaptive Inference in M-Estimation

[Student Paper Competition 1, (page 5)]
Kenta TAKATSU, *Carnegie Mellon University*
Arun Kumar KUCHIBHOTLA, *Carnegie Mellon University*

This manuscript studies a general approach to construct confidence sets for the solution of population-level optimization, commonly referred to as M-estimation. Statistical inference for M-estimation poses significant challenges due to the non-standard limiting behaviors of the corresponding estimator, which arise in settings with increasing dimension of parameters, non-smooth objectives, or constraints. We propose a simple and unified method that guarantees validity in both regular and irregular cases. Moreover, we provide a comprehensive width analysis of the proposed confidence set, showing that the convergence rate of the diameter is adaptive to the unknown degree of instance-specific regularity. We apply the proposed method to several high-dimensional and irregular statistical problems.

120. Micro-macro changepoint inference for periodic data sequences

[03.M2.I53, (page 29)]
Rebecca KILLICK, *Lancaster University / UC Santa Cruz*

When considering finer scale environmental data, e.g. daily or sub-daily, we have to model the finer scale periodicities and/or changes that become part of the 'noise' to the climate signal we wish to understand. However, it can be hard to disentangle the large scale, climate driven, changes amongst the finer scale changes. However, failure to model the finer behaviour can lead to incorrect inference on the large scale patterns. In addition, the finer scale data has larger, often nonstationary, second order behaviour than its monthly or yearly counterparts. To combat these issues of nonstationary second order structure and multiscale changes we propose a hierarchical periodic changepoint approach that separately models the within year (fine scale) changes from the across year (climate related) changes whilst allowing for a nonstationary error structure. We demonstrate the approach on temperature data and describe the interesting results.

121. SMART-MC: Sparse Matrix Estimation with Covariate-Based Transitions in Markov Chain Modeling of Multiple Sclerosis Disease Modifying Therapies

[Student Poster Competition, (page 7)]
Beomchang KIM, *Virginia Commonwealth University*
Priyam DAS, *Virginia Commonwealth University*
Zongqi XIA, *University of Pittsburgh*

A Markov model is a widely used tool for modeling sequences of events from a finite state-space and hence can be employed to identify the transition probabilities across treatments based on treatment sequence data. To understand how patientlevel covariates impact these treatment transitions, the transition probabilities are modeled as a function of patient covariates. This approach enables the visualization of the effect of patient-level covariates on the treatment transitions across patient visits. The proposed method automatically estimates the entries of the transition matrix with smaller numbers of empirical transitions as constant; the user can set desired cutoff of the number of empirical transition counts required for a particular transition probability to be estimated as a function of covariates. Firstly, this strategy automatically enforces the final estimated transition matrix to contain zeros at the locations corresponding to zero empirical transition counts, avoiding further complicated model constructs to handle sparsity, in an efficient manner. Secondly, it restricts estimation of transition probabilities as a function of covariates, when the number of empirical transitions is particularly small, thus avoiding the identifiability issue which might arise due to the $p > n$ scenario when estimating each transition probability as a function of patient covariates. To optimize the multi-modal likelihood, a parallelized scalable global optimization routine is also developed. The proposed method is applied to understand how the transitions across disease modifying therapies (DMTs) in Multiple Sclerosis (MS) patients are influenced by patient-level demographic and clinical phenotypes.

122. Wasserstein Geodesic Generator for Conditional Distributions

[01.E1.I13, (page 12)]
Younggeun KIM, *Michigan State University*

Generating samples given a specific label requires estimating conditional distributions. We propose a novel conditional generation framework characterized by Wasserstein spaces, a metric space of distributions defined by Wasserstein distances. Using optimal transport theory, we introduce the Wasserstein geodesic generator, a new conditional generator that learns the Wasserstein geodesic between conditional distributions of observed domains and optimal transport maps connecting them. This method also extends to estimating conditional distributions along unobserved intermediate domains on the Wasserstein

geodesic. Experiments on face images using lighting conditions as domain labels confirm the efficacy of our approach.

123 . Statistical-computational Tradeoffs for Recursive Adaptive Partitioning Estimators

[03.A1.I61, (page 32)]

Jason KLUSOWSKI, *Princeton University*
Yan Shuo TAN, *National University of Singapore*
Krishna BALASUBRAMANIAN, *University of California, Davis*

Recursive adaptive partitioning estimators, like decision trees and their ensembles, are effective for high-dimensional regression but usually rely on greedy training, which can become stuck at suboptimal solutions. We study this phenomenon in estimating sparse regression functions over binary features, showing that when the true function satisfies a certain structural property—Abbe et al. (2022)'s Merged Staircase Property (MSP)—greedy training achieves low estimation error with only a logarithmic number of samples in the feature count. In contrast, when MSP is absent, estimation becomes exponentially more difficult. Interestingly, this dichotomy between efficient and inefficient estimation resembles the behavior of two-layer neural networks trained with SGD in the mean-field regime. Meanwhile, ERM-trained recursive adaptive partitioning estimators achieve low estimation error with logarithmically many samples, regardless of MSP, revealing a fundamental statistical-computational trade-off for greedy training.

124. Adaptive Inference Techniques for Some Irregular Problems

[02.E1.I40, (page 23)]

Arun KUCHIBHOTLA, *Carnegie Mellon University*
Woonyoung CHANG, *Carnegie Mellon University*

In this talk, I will present construction of valid confidence sets for functional defined by a system of equations (e.g., linear regression, GLMs, any Z-estimation problem). We will discuss conditions for validity. Interestingly, some of the confidence sets presented are valid no matter the dimension of the parameter. I will also present some results regarding the diameter of the confidence set. Finally, I will provide some precise results in the context of linear regression with increasing dimension. This is joint work with Woonyoung Chang.

125 . Refining Clinical Trials: Calibrated Bayesian Extrapolation Using Data-Informed Priors

[02.A1.I36, (page 22)]

Maria KUDELA, *Pfizer*
Heliang SHI, *Pfizer*
Yuxi ZHAO, *Pfizer*
Margaret GAMALO, *Pfizer*

In tackling the challenges of patient recruitment and addressing ethical concerns about placebo controls in clinical trials, especially for rare diseases or pediatric groups, this study introduces a refined Bayesian framework to effectively integrate external data. The key to this approach lies in the calibration of data-driven priors, tailored to both baseline and outcome metrics, ensuring they meet scientific and regulatory standards. The methodology features a dynamic borrowing technique supported by a multi-layered calibration process, which hinges on the degree of similarity between the target population and external data sources, guided by the propensity overlap. The effectiveness of proposed approach was evaluated by simulation studies and a practical application. By emphasizing similarity in disease or response over variability in outcomes, this methodology offers a more ethical and precise alternative, improving upon conventional clinical trial practices.

126. Distribution Regression Using Conditional Deep Generative Models

[Student Poster Competition, (page 8)]

Shivam KUMAR, *University of Notre Dame*
Shivam KUMAR, *University of Notre Dame*
Yun YANG, *University of Maryland at College Park*
Lizhen LIN, *University of Maryland at College Park*

In this work, we explore the theoretical properties of conditional deep generative models under the statistical framework of distribution regression where the response variable lies in a high-dimensional ambient space but concentrates around a potentially lower-dimensional manifold. More specifically, we study the large-sample properties of a likelihood-based approach for estimating these models. Our results lead to the convergence rate of a sieve maximum likelihood estimator (MLE) for estimating the conditional distribution (and its devolved counterpart) of the response given predictors in the Hellinger (Wasserstein) metric. Our rates depend solely on the intrinsic dimension and smoothness of the true conditional distribution. These findings provide an explanation of why conditional deep generative models can circumvent the curse of dimensionality from the perspective of statistical foundations and demonstrate that they can learn a broader class of nearly

singular conditional distributions. Our analysis also emphasizes the importance of introducing a small noise perturbation to the data when they are supported sufficiently close to a manifold. Finally, in our numerical studies, we demonstrate the effective implementation of the proposed approach using both synthetic and real-world datasets, which also provide complementary validation to our theoretical findings.

127. F-modeling-Based empirical Bayes Estimator for Parameters in the Scale Family

[01.E1.I19, (page 14)]

YEIL KWON, *Wichita State University*

We address the problem of estimating multiple population variances from observed sample variances using an empirical Bayes framework. By assuming a general prior distribution for the variances, we develop various Bayes estimators under different loss functions. Notably, for one particular loss function, the Bayes estimator depends solely on the marginal cumulative distribution function (CDF) of the sample variances. By substituting this CDF with its empirical counterpart, we derive a practical empirical Bayes estimator, referred to as the F-modeling-based empirical Bayes estimator. We establish key theoretical properties of this estimator and showcase its performance through comprehensive simulation studies and applications to real-world data.

128 . Model-free dynamic treatment regimes with arbitrary number of treatments and stages

[02.M2.I26, (page 18)]

Nilanjana LAHA, *Texas A&M*
Nilson CHAPAGAIN, *Texas A&M*
Victoria CICHERSKI, *HEB*
Aaron SONABEND-W, *Google Research*

Patients suffering from chronic diseases may receive treatments at multiple time-points or stages. Our aim is to learn the best individualized treatment policy, or dynamic treatment regimes (DTR), in this longitudinal setting using existing patient data. It is known that the above optimization problem reduces to a sequential weighted classification problem. This paper considers simultaneously solving the above-mentioned classification problem through Fisher consistent surrogate losses, when the number of treatment-levels per stage can be arbitrary. Although computationally feasible Fisher consistent losses are available for special settings, e.g., the binary treatment case, a general theory of surrogate losses for the sequential DTR classification remained undeveloped. To address this, we establish necessary and sufficient conditions for DTR Fisher consistency within the class of non-negative, stagewise separable surrogate losses, which, to our knowledge, is the first result of its kind in the DTR literature. Furthermore, we show that many convex surrogate losses fail to be Fisher consistent for the DTR classification problem, and we formally establish this inconsistency for smooth, permutation equivariant, and relative-margin-based convex losses. Next, we use the sufficient conditions to construct smooth Fisher consistent surrogates. Since they are non-convex, the resulting simultaneous direct search method (SDSS) reduces to a smooth but non-convex optimization, for which we design a computationally fast, gradient-based algorithm. When the optimization error is small, we establish a sharp upper bound on the regret decay of SDSS for rich policy classes. We assess the numerical performance of SDSS through simulations. Finally, we demonstrate its real-world performance by applying it to estimate optimal fluid resuscitation strategies for severe septic patients using electronic health record data.

129 . Coherence-free Entrywise Estimation of Eigenvectors in Low-rank Signal-plus-noise Matrix Models

[03.E1.I67, (page 35)]

Keith LEVIN, *University of Wisconsin, Madison*

TBA

130 . Perspectives on Climate Model Evaluation

[Special Invited Session 1, (page 9)]

Bo LI, *Washington University in St. Louis*

Evaluating climate models by comparing their historical simulations to observed data is crucial for identifying systematic errors and guiding further model development. Climate models can be evaluated from various perspectives, each requiring specific methodologies tailored to the particular objectives. In this work, we present two examples of climate model evaluation. The first example assesses the marginal extreme behavior of a regional climate model against reanalysis data. We developed a multiple testing procedure based on the characteristics of spatial extremes to identify areas where the marginal extreme value distribution and return levels diverge. The second example focuses on ranking general circulation models of the Coupled Model Intercomparison Project (CMIP) by comparing the distribution of their daily temperature and precipitation data to ob-

servations. To achieve this, we introduced the Spherical Convolutional Wasserstein Distance (SCWD), which accounts for spatial variability through convolutional projections and quantifies local differences in the distribution of climate variables. We additionally applied SCWD to evaluate the progression from CMIP Phase 5 to Phase 6 in terms of their ability to produce realistic climatologies.

131 . Quantifying the Mediation Effect for Non-sparse High-dimensional Omics Mediators

[02.A1.I37, (page 22)]

Chunlin LI, *Iowa State University*

Understanding the mediating role of omics data is crucial for uncovering the biological mechanisms through which an established risk factor influences the onset or progression of a disease. When investigating the mediation pathways of cardiovascular outcomes in the multi-ethnic study of atherosclerosis (MESA) dataset, we found that many proteins likely have weak mediating effects that could collectively play a substantial mediating role. Despite extensive research on high-dimensional mediation analysis, existing methods have often fallen short in accurately quantifying the contribution of omics mediators, particularly those with weak effects. To address this issue, we propose a new, variance-based mediation analysis framework and develop flexible estimation and inferential procedures based on a mixed-effects working model. Through this innovative approach, we are able to accurately quantify the mediation effects and discover the weak effects that are largely mis-estimated by existing methods. We also examine the validity and effectiveness of the proposed methods from both theoretical and empirical perspectives. The proposed approach is general and complements the existing mediation analysis methodologies by offering new perspectives on the total and weak effects.

132 . SyNPar: Synthetic Null Parallelism for High-Power and Fast FDR Control in Feature Selection

[Special Invited Session 4, (page 28)]

Jingyi Jessica LI, *University of California, Los Angeles*
Changhu WANG, *University of California, Los Angeles*
Ziheng ZHANG, *University of California, Los Angeles*

Balancing FDR and statistical power to ensure reliable discoveries is a key challenge in high-dimensional feature selection. Although several FDR control methods have been proposed, most involve perturbing the original data, either by concatenating knockoff variables or splitting the data into two halves—both of which can lead to a loss of power. In this paper, we introduce a novel approach called Synthetic Null Parallelism (SyNPar), which controls the FDR in high-dimensional feature selection while preserving the original data. SyNPar generates synthetic null data from a model fitted to the original data and modified to reflect the null hypothesis. It then applies the same estimation procedure in parallel to both the original and synthetic null data to estimate coefficients that indicate feature importance. By comparing the coefficients estimated from the null data with those from the original data, SyNPar effectively identifies false positives, functioning as a numerical analog of a likelihood ratio test. We provide theoretical guarantees for FDR control at any desired level while ensuring that the power approaches one with high probability asymptotically. SyNPar is straightforward to implement and can be applied to a wide range of statistical models, including high-dimensional linear regression, generalized linear models, Cox models, and Gaussian graphical models. Through extensive simulations and real data applications, we demonstrate that SyNPar outperforms state-of-the-art methods, including knockoffs and data-splitting methods, in terms of FDR control, power, and computational efficiency.

133 . Estimating the Global Average Treatment Effect under Structured Interference

[02.M1.I22, (page 15)]

Shuangning LI, *University of Chicago*
Kevin HAN, *Meta*
Johan UGANDER, *Stanford University*

The field of causal inference develops methods for estimating treatment effects, often relying on the Stable Unit Treatment Value Assumption (SUTVA), which states that a unit's outcome depends only on its own treatment. However, in many real-world settings, SUTVA is violated due to interference—where the treatment assigned to one unit influences the outcomes of others. Such interference can arise from social interactions among units or competition for shared resources, complicating causal analysis and leading to biased estimates. Fortunately, in many cases, interference follows structured patterns that can potentially be leveraged for more accurate estimation. In this paper, we examine and formalize two specific forms of structured interference—monotone interference and submodular interference—which we believe arise in many practical settings. We investigate how incorporating these structures can improve

causal effect estimation. Our main contributions are (i) a set of bounds relating key interference estimands under these structural assumptions and (ii) new estimators that integrate these structures through constrained optimization. Since these constraints may introduce bias, we further develop debiasing techniques based on treatment regeneration and bootstrap methods to mitigate this issue.

134. Cluster-robust inference with a single treated cluster

[02.A1.I32, (page 20)]

Xinran LI, *University of Chicago*
Chun Pong LAU, *University of Chicago*
Xinran LI, *University of Chicago*

The paper considers inference when there is a single treated cluster and a fixed number of control clusters. This setting is common in empirical work, especially in difference-in-differences designs. We use the t-statistic and develop suitable critical values to conduct valid inference under weak assumptions allowing for unknown dependence within clusters. The proposed test works for any significance level when there are at least two control clusters. For most conventional significance levels and numbers of clusters, our critical values can be easily computed without any optimization. In other cases, one-dimensional numerical optimization is needed and is often computationally efficient; we have also tabulated common critical values in the paper so researchers can use them readily. Moreover, our procedure does not involve estimating the variances. It only requires specifying the relative heterogeneity between the variances from the treated cluster and some, but not necessarily all, control clusters. We illustrate our method in simulations and empirical applications.

135. Advancing Clinical Dementia Rating (CDR) Analysis: Efficiency Gains Through Item Response Theory

[03.A1.I57, (page 31)]

Yan LI, *Washington University in St. Louis*

Background CDR sum of boxes (CDR-SB) is widely used as a primary outcome in Alzheimer disease (AD) research due to its reliability and accuracy. However, it is an ordinal rank measure instead of a continuous measure and therefore may not be able to capture subtle changes in cognition. Additionally, it weights each domain equally, without accounting for the patterns of the responses scores across CDR items. To generate a more optimal score, we developed a bi-factor model based on the item response theory (IRT) which can account for different response pattern, item difficulty and discrimination level to provide an improved estimate of the dementia severity. We demonstrated the superiority of the IRT score in two independent cohorts. Methods Baseline item level CDR data from 2949 participants enrolled in the Washington University Knight ADRC study were used to develop the IRT model. We compared the fit of four different models: (1) Model A: a unidimensional IRT model with all items contributing to a general factor; (2) Model B: a multidimensional IRT model with six correlated factors for six domains in the CDR; (3) Model C: a bi-factor model with a general factor contributed by all items, and six uncorrelated factors corresponding to the six domains of the CDR; (4) Model D: the same bi-factor model as Model C but allowing for correlations between the domain specific factors. The best fitting IRT model was then applied to the longitudinal item-level CDR data to generate the IRT score (a continuous measure) for each participant at each visit. Results Correlations between the CDR IRT score and SB were consistent at different visits (0.92 0.94). The IRT score had a much larger effect size ($>$ 40%) than that of CDR-SB in the difference of rate of change when comparing progressor to stable cognitively normal participants. For the CDR-SB 0 group, the IRT score was significantly associated with some of the clinical/cognitive outcomes and biomarkers. For the CDR-SB $>$0 group, the IRT score had slightly higher correlations with other outcomes than CDR-SB. Using IRT score as the primary outcome in clinical trials could save 10% 30% of the sample size compared to using CDR-SB. Conclusion The CDR IRT score could potentially be used as a more efficient outcome in AD research to gain precision in detecting cognitive decline and improve power (or reduce sample size).

136 . Quantifying common and distinct information in multiomic single-cell data

[03.M2.I56, (page 30)]

Kevin LIN, *University of Washington*
Kevin LIN, *University of Washington*
Nancy ZHANG, *University of Pennsylvania*
Haoye YANG, *University of Chicago*

Recently, multi-modal single-cell data has been growing in popularity and provides new opportunities to learn how different modalities coordinate within each cell. Many existing dimension reduction methods for such data estimate a low-dimensional embedding that captures all the axes of variation from either modality. While these current methods are useful,

we develop the Tilted-CCA in this talk to perform a fundamentally different task. This method is a novel matrix factorization that estimates low-dimensional embeddings separating the axes of variation shared between both modalities (i.e., "common geometry," capturing the coordination between both modalities) from axes of variation unique to a particular modality (i.e., "distinct geometry"). Methodologically, Tilted-CCA achieves this by combining ideas from Canonical Correlation Analysis (CCA) and density clustering. Our method first uses the nearest-neighbor graphs from each modality to infer the common geometry between both modalities and decomposes the canonical scores from CCA to approximate this geometry. Biologically, Tilted-CCA unveils the cellular dynamics in developmental systems based on the proportion of variation between the common and distinct embeddings. With any remaining time, I'll discuss recent extensions my lab has been working on to use these same statistical principles but within a deep-learning framework to learn more complex relations..

137. Structure Learning and Statistical Inference in Tensor Ising Models

[Student Poster Competition, (page 8)]

Tianyu LIU, *National University of Singapore*
Tianyu LIU, *Department of Statistics and Data Science, National University of Singapore*
Somabha MUKHERJEE, *Department of Statistics and Data Science, National University of Singapore*
Bhaswar BHATTACHARYA, *Department of Statistics, University of Pennsylvania*

The Ising model is a discrete exponential family for modeling dependent binary data. The classical Ising model was initially used by physicists as a framework for ferromagnetism, where the sufficient statistic consists of quadratic terms designed to capture the pairwise interactions between the variables sitting on the nodes of a network. However, in many applications, the dependencies can be more complex in nature and may involve multi-body interactions, e.g. peer group effects in social science. A useful variant of the Ising model that captures these effects of the peer group is the k-tensor Ising model, where the sufficient statistic consists of a multilinear form of degree k. In this presentation, I will address two main areas of research in the field of tensor Ising models, namely structure learning (inferring the entire tensor) and parameter estimation, based on observations from the underlying tensor Ising model. For structure learning, the results are centered around establishing theoretical guarantees of some useful competing structure recovery algorithms, such as the "pseudo-likelihood based node-wise LASSO" and the "interaction screening". It is shown that both approaches, with proper regularization, retrieve the underlying hyper-network structure using a sample size logarithmic in the number of network nodes, and exponential in the maximum interaction strength and maximum node-degree. For parameter estimation, the Cramér-type moderate deviation for magnetization in p-spin Curie-Weiss models is proved and the Berry-Esseen bounds for magnetization in this model are derived. In particular, a $N^{-1/4}$ rate of convergence is obtained at all parameters where the limiting distribution of the magnetization is non-Gaussian.

138 . Empirical Error Estimates for Graph Sparsification

[03.E1.I64, (page 34)]

Miles LOPES, *UC Davis*
Siyao WANG,

Graph sparsification is a well-established technique for accelerating graph-based learning algorithms, which uses edge sampling to approximate dense graphs with sparse ones. Because the sparsification error is random and unknown, users must contend with uncertainty about the reliability of downstream computations. Although it is possible for users to obtain conceptual guidance from theoretical error bounds in the literature, such results are typically impractical at a numerical level. Taking an alternative approach, we propose to address these issues from a data-driven perspective by computing empirical error estimates. The proposed error estimates are highly versatile, and we demonstrate this in four use cases: Laplacian matrix approximation, graph cut queries, graph-structured regression, and spectral clustering. Moreover, we provide two theoretical guarantees for the error estimates, and explain why the cost of computing them is manageable in comparison to the overall cost of a typical graph sparsification workflow. (Joint work with Siyao Wang.)

139 . Computationally efficient reductions between some statistical models

[Student Poster Competition, (page 8)]

Mengqi LOU, *Georgia Institute of Technology*
Guy BRESLER, *MIT*
Ashwin PANANJADY, *Georgia Institute of Technology*

We study the problem of approximately transforming a sample from a source statistical model to a sample from a target statistical model without know-

ing the parameters of the source model, and construct several computationally efficient such reductions between canonical statistical experiments. In particular, we provide computationally efficient procedures that approximately reduce uniform, Erlang, and Laplace location models to general target families. We illustrate our methodology by establishing nonasymptotic reductions between some canonical high-dimensional problems, spanning mixtures of experts, phase retrieval, and signal denoising. Notably, the reductions are structure-preserving and can accommodate missing data. We also point to a possible application in transforming one differentially private mechanism to another.

140. Generalized Bayesian Inference for Dynamic Random Dot Product Graphs

[03.M1.I49, (page 27)]

Joshua LOYAL, *Florida State University*
Joshua LOYAL, *Florida State University*

The random dot product graph is a popular model for network data with extensions that accommodate dynamic (time-varying) networks. However, two significant deficiencies exist in the dynamic random dot product graph literature: (1) no coherent Bayesian way to update one's prior beliefs about the model parameters due to their complicated constraints, and (2) no approach to forecast future networks with meaningful uncertainty quantification. This work proposes a generalized Bayesian framework that addresses these needs using a Gibbs posterior that represents a coherent updating of Bayesian beliefs based on a least-squares loss function. Furthermore, we establish the consistency and contraction rate of this Gibbs posterior under commonly adopted Gaussian random walk priors. For estimation, we develop a fast Gibbs sampler with a time complexity that is linear in both the number of time points and observed edges in the dynamic network. Simulations and real data analyses show that the proposed method's in-sample and forecasting performance outperforms that of competitors.

141. A Bayesian Hybrid Phase 2 Design Incorporating Historical Monotherapy Data and Covariate Adjustment

[03.M1.I50, (page 27)]

Zhaohua LU, *Daiichi-Sankyo Inc.*
Yiyuan HUANG, *University of Michigan*
Philip HE, *Daiichi-Sankyo*

In the realm of early-phase drug development, evaluating the added efficacy of a new agent to a monotherapy is crucial. A Bayesian hybrid design that dynamically borrows from historical monotherapy data presents a powerful method to enhance the efficiency and decision-making process of studies. Lu et al. (2024) introduced a dynamic borrowing framework for binary outcomes, utilizing the dynamic power prior (DPP) and assessing the outcome similarity between the study control and historical control groups. This method proves to be more robust compared to traditional single-arm designs, significantly enhancing statistical power during the design phase. Additionally, incorporating key covariates at the analysis stage refines the modeling by mitigating variability and confounding effects. In this research, we propose an analysis framework that integrates covariates through propensity scores into the DPP (PS-DPP), facilitating more informed borrowing by considering the similarity between historical and current control data in terms of both covariates and outcomes. Our simulation studies demonstrate that PS-DPP enhances analytical performance, particularly when there are notable differences in response rates and covariates between historical and current controls.

142. Conformal Prediction for Dyadic Regression under Structured Missingness

[01.A1.I8, (page 10)]

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

Dyadic regression, which involves modeling a relational matrix given covariate information, is an important task in statistical network analysis. We consider uncertainty quantification for dyadic regression models using conformal prediction. We establish finite-sample validity of our procedures for various sampling mechanisms under a joint exchangeability assumption. We also show that, under certain conditions, it is possible to construct asymptotically valid prediction intervals for a missing entry under a structured missingness assumption.

143. Distribution-free Inference for Model Class Risk

[02.A1.I35, (page 21)]

Yuetian LUO, *University of Chicago*
Manuel MULLER, *University of Cambridge*
Rina FOYGEL BARBER, *University of Chicago*

In statistics and machine learning, when we train a fitted model on available data, we typically want

to ensure that we are searching within a model class that contains at least one accurate model—that is, we would like to ensure an upper bound on the model class risk (the lowest possible risk that can be attained by any model in the class). However, it is also of interest to establish lower bounds on the model class risk, for instance so that we can determine whether our fitted model is at least approximately optimal within the class, or, so that we can decide whether the model class is unsuitable for the particular task at hand. Particularly in the setting of interpolation learning where machine learning models are trained to reach zero error on the training data, we might ask if, at the very least, a positive lower bound on the model class risk is possible—or are we unable to detect that "all models are wrong"? In this talk, we aim to address these questions in a distribution-free setting by establishing a model-agnostic, fundamental hardness result for the problem of constructing a lower bound on the best test error achievable over a model class, and examine its implications on specific model classes such as tree-based methods and linear regression.

144. MiLC for Adjusting of Compositionality and Unobserved Confounding in Microbiome Data

[03.M1.I52, (page 28)]

Siyuan MA, *Vanderbilt University Medical Center*
Chih-Ting YANG, *Vanderbilt University Medical Center*
Yu SHYR, *Vanderbilt University Medical Center*
Chris MCKENNAN, *University of Pittsburgh*

Recent research has raised concerns over false discoveries in microbiome studies, in particular, regarding popular differential abundance (DA) analysis. Relatedly, statistical method development focused on the data's compositionality as a cause for false discoveries. We examine another, potentially stronger cause: unobserved microbiome-wide confounding (e.g., population heterogeneity, uncollected host covariates). Using real-data evidence, we show that unobserved confounding inflates false discoveries in microbiome DA, even more than data compositionality. To address this, we develop a novel factor-modeling regression method (Microbiome Latent Confounder DA, or MiLC) to estimate unobserved confounding factors and control false discoveries. MiLC can be applied to both relative abundance and read count microbiome data. We validate its performance to control false discoveries in comparison with existing methods with extensive simulation- and real-data-based benchmarking. Our work provides one of the first systematic efforts to evaluate and correct for unobserved confounding in microbiome DA, improving the reliability of microbiome research findings.

145. AI-powered Clinical Development: Integrating Clinical Insights with Statistical Innovations

[01.A1.I5, (page 9)]

Will MA, *HopeAI*

While Large Language Models excel at general knowledge, clinical development demands a more nuanced approach that combines domain expertise with statistical rigor. We present a novel dual-agent system consisting of an AI Clinical Scientist that synthesizes complex clinical evidence, working in tandem with an AI Statistician that implements real-time statistical methods. This AI-augmented approach shows promise for accelerating and strengthening clinical trial design, providing clinical development teams with powerful new tools for evidence-based decision-making.

146. Confidence Interval Construction and Conditional Variance Estimation with Dense ReLU Networks

[02.A1.I34, (page 21)]

Carlos Misael MADRID PADILLA, *Washington University in St Louis*
Oscar Hernan MADRID PADILLA,
Yik Lun KEI,
Yanzhen CHEN,

In this talk, I will discuss recent progress on conditional variance estimation and confidence interval construction in nonparametric regression using deep neural networks with ReLU activation. I will present a residual-based framework for estimating conditional variances, along with non-asymptotic error bounds that hold under both homoscedastic and heteroscedastic noise structures. I will also introduce a robust ReLU-based bootstrap procedure for constructing confidence intervals with provable coverage guarantees. This approach offers a practical and theoretically sound framework for uncertainty quantification in deep learning-based regression.

147. Network two-sample test for block models

[02.M1.I21, (page 15)]

Oscar Hernan MADRID PADILLA, *University of California, Los Angeles*
Chung Kyong NGUEN, *University of California, Los Angeles*

Arash AMINI,

We consider the two-sample testing problem for networks, where the goal is to determine whether two sets of networks originated from the same stochastic model. Assuming no vertex correspondence and allowing for different numbers of nodes, we address a fundamental network testing problem that goes beyond simple adjacency matrix comparisons. We adopt the stochastic block model (SBM) for network distributions, due to their interpretability and the potential to approximate more general models. The lack of meaningful node labels and vertex correspondence translate to a graph matching challenge when developing a test for SBMs. We introduce an efficient algorithm to match estimated network parameters, allowing us to properly combine and contrast information within and across samples, leading to a powerful test. We show that the matching algorithm, and the overall test are consistent, under mild conditions on the sparsity of the networks and the sample sizes, and derive a chi-squared asymptotic null distribution for the test. Through a mixture of theoretical insights and empirical validations, including experiments with both synthetic and real-world data, this study advances robust statistical inference for complex network data.

148 . Long-range competition on the torus

[01.E1.I16, (page 13)]

Neeladri MAITRA, *University of Illinois at Urbana-Champaign*
Bas LODEWIJKS,

In this talk we look at a competition process between two growth models with long-range correlations on the torus T_n^d of size n in dimension d. We append the edge set of the torus T_n^d by including all non-nearest-neighbour edges, and from two source vertices, two first-passage percolation (FPP) processes start flowing on T_n^d and compete to cover the sites, where the cost of traversing an edge is increasing with respect to the distance between the vertices it connects. Under a *mean-field* assumption, we study *coexistence*, the event that both types reach a positive proportion of the torus, and identify precisely when coexistence occurs. In the case of absence of coexistence, we outline several phase transitions in the size of the losing type, depending on the relation between the local spreading rates of both the types.

149. Tensor-on-Tensor Times Series Regression for Integrated One-step Analysis of fMRI Data

[02.M1.I21, (page 15)]

Ranjan MAITRA, *Iowa State University*
Subrata PAL,

Data acquisition in a functional Magnetic Resonance Imaging (fMRI) activation detection experiment yields a massively structured array- or tensor-variate dataset that need to be analyzed with respect to a set of time-varying stimuli and possibly other covariates. The conventional approach employs a two-stage analysis: The first stage fits an univariate regression on the time series data at each individual voxel and reduces the voxel-wise data to a single statistic. The statistical parametric map formed from these voxel-wise test statistics is then fed into a second-stage analyis that potentially incorporates spatial context between the voxels and identifies activation within them. We develop holistic yet practical tensor-variate methodology that provides one-stage tensor-variate regression modeling of the entire time series array-variate dataset. Low-rank specifications on the tensor-variate regression parameters and Kronecker separable error covariance tensors make our innovation feasible. A block relaxation algorithm provides maximum likelihood estimates of the model parameters. A R package, with C backends for computational feasibility, operationalizes our methods. Performance on different real-data-imitating simulation studies and a functional MRI study about Major Depressive Disorder demonstrate the stability of our approach and that it can reliably identify cerebral regions that are significantly activated.

150 . Measuring and Ensuring Consistency in Language Models

[02.M1.I24, (page 17)]

Subho MAJUMDAR, *Vijil*
Harsh RAJ, *Northeastern University*
Domenic ROSATI, *Dalhousie University*
Vipul GUPTA, *Penn State University*

Despite their impressive capabilities, language models often produce inconsistent outputs when presented with semantically equivalent inputs, undermining their reliability in critical applications. Our research explores methodologies for measuring this inconsistency problem and developing effective solutions. We propose semantic consistency metrics to evaluate multiple consistency measures across diverse model architectures and sizes—finding that lack of consistency is a problem prevalent in today's language models. In addition, we implement a guided prompting technique called Chain-of-Guidance (CoG). CoG shows significant consistency

improvements for closed-book question-answering tasks across multiple experimental scenarios. We conclude by outlining future research areas that are promising directions of statistical inquiry. These include robust training and loss landscape analysis—advancements critical for deploying language models where predictability and reliability are essential.

151 . Bayesian Joint Additive Factor Models for Multiview Learning

[03.A1.I59, (page 31)]

Himel MALLICK, *Cornell University*
David DUNSON, *Duke University*
Niccolo ANCESCHI, *Duke University*
Federico FERRARI, *Merck Research Laboratories*

It is increasingly common in a wide variety of applied settings to collect data of multiple different types on the same set of samples. Our particular focus in this article is on studying relationships between such multiview features and responses. A motivating application arises in the context of precision medicine where multi-omics data are collected to correlate with clinical outcomes. It is of interest to infer dependence within and across views while combining multimodal information to improve the prediction of outcomes. The signal-to-noise ratio can vary substantially across views, motivating more nuanced statistical tools beyond standard late and early fusion. This challenge comes with the need to preserve interpretability, select features, and obtain accurate uncertainty quantification. We propose a joint additive factor regression model (JAFAR) with a structured additive design, accounting for shared and view-specific components. We ensure identifiability via a novel dependent cumulative shrinkage process (D-CUSP) prior. We provide an efficient implementation via a partially collapsed Gibbs sampler and extend our approach to allow flexible feature and outcome distributions. Prediction of time-to-labor onset from immunome, metabolome, and proteome data illustrates performance gains against state-of-the-art competitors. Our open-source software (R package) is publicly available.

152 . Bayesian Models for Joint Selection of Features and Auto-Regressive Lags: Theory and Applications in Environmental and Financial Forecasting

[Student Poster Competition, (page 8)]

Alokesh MANNA, *University of Connecticut*
Sujit GHOSH,

This paper presents Bayesian variable selection methods for linear regression models with autocorrelated errors, addressing the simultaneous selection of predictor variables (including their lagged values) and lagged error terms. Such models have broad applications in fields like finance (e.g., stock price log-returns), environmental science (e.g., weather forecasting and water table dynamics), and other disciplines where autocorrelation plays a key role. The inclusion of lagged exogenous variables enables the modeling of delayed or prolonged effects of external factors on the response variable. We propose a two-step Markov Chain Monte Carlo (MCMC) procedure utilizing spike-and-slab priors to perform variable selection and prediction. The methodology ensures posterior selection consistency for marginal inclusion probabilities under suitable regularity conditions, even when the number of covariates grows exponentially with the number of time points. Extensive numerical experiments using simulated data and real-world applications, including predictions of water table depth and the S&P 500 stock market, demonstrate the effectiveness of the proposed approach. The results highlight its superior variable selection accuracy, consistency, and predictive performance, characterized by low mean squared error (MSE) and other performance measures. This framework provides a robust variable selection and forecasting tool in autoregressive time series models.

153 . Exploring potential treatment effect heterogeneities in clinical trials with time to event endpoints

[01.M2.I4, (page 4)]

Alejandro MANTERO, *GSK*

AI/ML represents a body of research that includes methods for finding more complex associations between measurements than what is traditionally done. Thus far, AI/ML has seen implementation in many respects related to pharmaceutical R&D (Research & Development), such as clinical operations, drug discovery, and patient selection, but it has been lacking in research pertaining directly to clinical trial design and analysis [Gamalo, 2021, Weissler 2021]. Increased clinical data sharing effort, and availability of real-world data sources (RWD) creates a space where AI/ML could add valuable- insight to clinical trial design. In this paper, we present a case study in NSCLC (Non-Small Cell Lung Cancer) with a time to event endpoint, OS (overall survival), using the sidClustering (Mantero, Stat Anal Data Min. 2020) ML and K-prototypes (an extension to K-means for mixed continuous/categorical data) methods. Man-

tero et al. (Stat Anal Data Min. 2020) introduced the sidClustering method, a new random forest-based approach for unsupervised machine learning, which finds clusters of observations based on measured features (i.e., independent variables). We used patient baseline characteristics to determine areas of heterogeneous treatment effect on OS. The objective of this work was to discover clusters of patient profiles using their baseline characteristics who respond differentially to treatment. The baseline characteristics were clustered while holding out the survival outcome to serve as external validation of the clusters. These discovered patient subgroups can be used to inform future clinical trials with better population targeting and potentially resulting in smaller required sample sizes and better patient outcomes. Through this case study, we illustrate common patterns and how a future trial might be designed based on this information.

154 . Genetic fine mapping of high-dimensional traits, with application to metabolite genome-wide association studies

[02.M2.I28, (page 19)]

Chris MCKENNAN, *University of Pittsburgh*
Weiqiong HUANG, *University of Pittsburgh*
Emily HECTOR, *North Carolina State University*

Multi-trait genetic fine mapping seeks to determine the causal impact of genetic variants on several traits at once, thereby facilitating investigation into the shared genetic origin (pleiotropy) of traits. However, existing methods to perform multi-trait fine mapping can only consider a small number of traits, which precludes the analysis of high dimensional 'omic' and biobank data. To address this, we develop a novel model, theoretical framework, and set of Bayesian methods capable of simultaneously fine mapping hundreds to thousands of traits using only genome-wide association study (GWAS) summary statistics. We apply our method to metabolic traits, where we develop nonparametric priors for genetic effects on metabolite levels that use metabolic pathway information to enable interpretable analyses at the pathway level. Our analysis of real metabolite GWAS data shows our method is sensitive enough to confirm results from a study with twenty times our sample size, and can make inferences that would otherwise be impossible with existing methods.

155 . LatticeVision: Image-to-Image Networks for Modeling Non-Stationary Spatial Data

[02.A1.I33, (page 21)]

Daniel MCKENZIE, *Colorado School of Mines*
Antony SIKORSKI, *Colorado School of Mines*
Michael IVANITSKIY, *Colorado School of Mines*
Doug NYCHKA, *Colorado School of Mines*

When fitting a statistical model to observational data, parameters are usually inferred by methods such as maximum likelihood estimation (MLE). However for large, non-stationary models commonly used in spatial statistics, MLE becomes intractable. Consequently, a recent line of work known as _neural amortized inference_ has investigated using neural networks to approximate the function mapping observations to model parameters directly. This avoids optimizing, or even computing, the likelihood function.

We further this line of work by observing that for many spatial autoregressive (SAR) models, the parameters may be arranged in a regular grid resembling the pixels of an image. Thus, image-to-image networks originally introduced for tasks such as image segmentation may be used to estimate these spatially varying parameters. We demonstrate the power of our approach by fitting a non-stationary SAR model to global climate model (GCM) output.

156 . Large scale few-shot learning through parameter pooling

[02.A1.I31, (page 20)]

Semhar MICHAEL, *South Dakota State University*
Andrew SIMPSON, *South Dakota State University*

In Large-scale classification problems with few shots, the data is characterized by the observations partitioned into many classes with a limited number of samples per class. This brings a significant challenge to classical probabilistic and machine learning approaches. Given the nature of the few-shot learning framework, classical statistical methods make strong assumptions about the data-generating process. In many cases, to get non-singular and stable estimates for the covariance matrices of each class, it is often assumed that each class has the same covariance matrix as in linear discriminant analysis (LDA). In this framework, given that the number of classes tends towards infinity, this assumption is extreme. The strong assumption of LDA is relaxed, and stable estimates of the covariance matrices are obtained. This talk will focus on our proposed mixture model of covariance matrices for this classification problem. Using simulation studies will demonstrate the utility of the proposed methodology. The application of the method to the forensic source classification problem

will be discussed.

157. Half-orthogonalized Neural Network for Estimating Yield Response Function Using On-farm Experiment Data

[04.M1.I69, (page 36)]

Taro MIENO, *University of Nebraska Lincoln*
Mona MOISAVI, *University of Nebraska Lincoln*
David BULLOCK, *University of Illinois at Urbana Champaign*

Various machine learning (ML) models have been applied to estimate yield response function using data obtained from on-farm field trials. This paper proposes a new approach termed "Half-orthogonalized" neural network and examine its performance compared to popular ML methods including neural network and random forest. Borrowing from the idea of double machine learning approaches for causal inference, our approach orthogonalize the dependent variable (but not the independent variables) and then applies neural network approach on the orthogonalized dependent variable. Via Monte Carlo simulation, we show that this approach far outperform the other models.

158. Multistage drop-the-losers designs for selecting the effective treatment(s) and estimating their worth

[01.M2.I2, (page 4)]

Neeraj MISRA, *Indian Institute of Technology Kanpur*
Yogesh KATARIYA, *Indian Institute of Technology Kanpur*

Consider k (≥ 2) treatments whose effects are described by independent Gaussian responses with unknown means and a common variance. For the purpose of selecting effective treatments (drugs) and estimating their worth (defined as the average of mean responses of selected treatments), we will discuss two multistage drop-the-losers designs (DLDs). Since the bias of estimators is crucial in clinical studies, we will discuss the Uniformly Minimum Variance Conditionally Unbiased Estimation (UMVCUE) of the worth of the selected treatments, conditioned on the indices of treatments selected. The mean squared error (MSE) and the bias performances of the UMVCUE against those of the maximum likelihood estimator will also be discussed.

159. Bayesian inference for COVID-19 transmission dynamics using a modified SEIR model

[04.M2.I73, (page 37)]

Anirban MONDAL, *Case Western Reserve University*
Kai YIN, *Case Western Reserve University*
Paromita BANERJEE, *John Carroll University*
David GURARIE, *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.

160. Dimension Reduction for Regression Model with Spatially Correlated Data

[02.A1.I33, (page 21)]

Hossein MORADI REKABDARKOLAEE, *South Dakota State University*

Natural sciences such as geology and forestry often utilize regression models for spatial data with high-dimensional predictors and moderate sample sizes. In this case, efficient estimation of the regression parameters is crucial for both model interpretation and prediction. The predictor envelope is a method of dimension reduction for linear regression with multivariate predictors that assumes certain linear combinations of the predictors are immaterial to the regression. The method can result in substantial gains in estimation efficiency and prediction accuracy over traditional maximum likelihood and least squares estimates. While predictor envelopes have been developed and studied for independent data, no work has been done adapting predictor envelopes to

spatial data. In this work, the predictor envelope is adapted to a popular spatial model to form the spatial predictor envelope (SPE). Maximum likelihood estimates for the SPE are derived, along with asymptotic distributions for the estimates given certain assumptions, showing the SPE estimates to be asymptotically more efficient than generalized least squares, the typical spatial regression estimates. Further, we study the SPE in the context of spatial prediction, or universal kriging, discussing the contexts in which the SPE can provide gains over the typical universal kriging predictions. The effectiveness of the proposed model is illustrated through simulation studies and the analysis of a geochemical data set, predicting rare earth element concentrations within an oil and gas reserve in Wyoming.

161. Model-based inference for multiple dependent graphs – Riten Mitra's contributions

[02.M1.I20, (page 15)]

Peter MUELLER, *Professor, Department of Statistics and Data Sciences, Department of Mathematics, UT Austin*

TBA

162. TBD

[01.A1.I7, (page 10)]

Debarghya MUKHERJEE, *Boston University*

TBA

163 . Network-linked high-dimensional multinomial Probit

[02.E1.I38, (page 22)]

Gourab MUKHERJEE, *University of Southern California*
Rashmi BHUYIAN, *USC*
Adel JAVANMARD, *USC*

The multinomial probit model (MNP) is widely used for analyzing unordered categorical data. In a host of contemporary applications, we encounter cross-sectional datasets with categorical responses and high-dimensional covariates. In absence of repeated observations from the respondents, MNP models in these applications are often equipped with additional network structures that leverage contiguity in the responses of similar units. Pooling information across similar units through these network structures can provide significantly better inference in these data scarcity problems. However, estimating the effects of sparse high-dimensional covariates in the presence of network linkages among the responses is challenging. We develop a composite likelihood based method that consistently estimates the covariate effects. We provide decision theoretic guarantees on the operational characteristic of the proposed algorithm. We demonstrate the application of the proposed method in spatial autocorrelation network structured MNP models.

164. Method-of-Moments Inference for GLMs and Doubly Robust Functionals under Proportional Asymptotics

[04.M1.I70, (page 36)]

Rajarshi MUKHERJEE, *Harvard University*
Xingyu CHEN, *Shanghai Jiao Tong University*
Liu LIN , *Shanghai Jiao Tong University*

In this paper, we consider the estimation of regression coefficients and signal-to-noise (SNR) ratio in high-dimensional Generalized Linear Models (GLMs), and explore their implications in inferring popular estimands such as average treatment effects in high dimensional observational studies. Under the "proportional asymptotic" regime and Gaussian covariates with known (population) covariance Sigma we derive root-n-Consistent and Asymptotically Normal (CAN) estimators of our targets of inference through a Method-of-Moments type of estimators that bypasses estimation of high dimensional nuisance functions and hyperparameter tuning altogether. Additionally, under non-Gaussian covariates, we demonstrate universality of our results under certain additional assumptions on the regression coefficients and Sigma . We also demonstrate that knowing Sigma can be relaxed in our proposed methodology. Finally, we complement our theoretical results with extensive numerical experiments, in comparisons with competing methods.

165. Mixing Phases and Metastability of the Glauber Dynamics in Tensor Curie-Weiss Models

[01.A1.I8, (page 10)]

Somabha MUKHERJEE, *National University of Singapore*
Ramkrishna SAMANTA, *University College London*
Jiang ZHANG, *National University of Singapore*

In this talk, we consider the p-spin generalization of the Curie-Weiss Ising model with an external field, and identify three disjoint regions partitioning the parameter space, where the corresponding Glauber dynamics exhibits three different orders of mixing times. The construction of these disjoint regions depends on the number of local maximizers of a certain negative free-energy function H, and the behavior of

the second derivative of H at such a local maximizer. We provide an explicit description of the geometry of these three different phases within the parameter space. Finally, we show that if H has multiple local maximizers (metastable states) which slow down the mixing of the original Glauber dynamics exponentially, then one can still create a restricted version of the original Glauber dynamics, which mixes fast. This is a joint work with Ramkrishna Samanta and Jiang Zhang.

166. CLT in HD Bayesian linear regression

[01.E1.I18, (page 13)]

Sumit MUKHERJEE, *Columbia University*
Seunghyun LEE, *Columbia University*
Nabarun DEB, *University of Chicago*

In this talk we study a CLT for linear statistics of the posterior in high dimensional Bayesian linear regression with an iid prior. In contrast to the existing literature which focuses on the high SNR regime where the prior washes away and one obtains Bernstein-von-Mises type results, our work focuses on the low SNR regime where the prior has a significant effect. As application of our result, we derive asymptotic coverage of posterior credible intervals when the prior is mis-specified.

167. Bayesian Analysis of Space Sustainability issues

[04.M1.I69, (page 36)]

Ujjal MUKHERJEE, *University of Illinois Urbana Champaign*
Ujjal MUKHERJEE, *University of Illinois Urbana-Champaign*
Snigdhansu CHATTERJEE, *University of Maryland Baltimore County*

Space launches, particularly in low earth orbits, has been increasing significantly in recent years. At the same time, the average expected life of each satellite has been decreasing. Space agencies like NASA and ESA have been considering measures for sustainably exploring space without major concerns of unaccounted space debris re-entering Earth atmosphere. In this paper, we collected a unique dataset from NASA and ESA databases to conduct a Bayesian analysis to forecast the demand for space satellites and the extent of sustainability concerns that space debris can cause to Earth's atmosphere.

168. Data Science Problems in the Tech Industry

[03.M1.I47, (page 26)]

Jami MULGRAVE, *North Carolina State University*

Measurement is a key aspect of the tech industry. Measurement involves assessing various aspects to track progress, identify areas for improvement, and ensure alignment with business goals. This includes measuring employee productivity, technical performance, and the impact of technology changes on business operations and customer experience. In this talk, I will discuss data science issues that arise in the area of measurement.

169. TBD

[02.A1.I35, (page 21)]

Vidya MUTHUKUMAR, *Georgia Tech*

TBA

170 . Estimating stationary mass, frequency by frequency

[Student Poster Competition, (page 8)]

Milind NAKUL, *Georgia Institute of Technology*
Vidya MUTHUKUMAR, *Georgia Institute of Technology*
Ashwin PANANJADY, *Georgia Institute of Technology*

Suppose we observe a trajectory of length n from an α-mixing stochastic process over a finite but potentially large state space. We consider the problem of estimating the probability mass placed by the stationary distribution of any such process on elements that occur with a certain frequency in the observed sequence. We estimate this vector of probabilities in total variation distance, showing universal consistency in n and recovering known results for i.i.d. sequences as special cases. Our proposed methodology carefully combines the plug-in (or empirical) estimator with a recently-proposed modification of the Good–Turing estimator called WingIt, which was originally developed for Markovian sequences. En route to controlling the error of our estimator, we develop new performance bounds on WingIt and the plug-in estimator for α-mixing stochastic processes. Importantly, the extensively used method of Poissonization can no longer be applied in our non i.i.d. setting, and so we develop complementary tools—including concentration inequalities for a natural self-normalized statistic of mixing sequences—that may prove independently useful in the design and analysis of estimators for related problems.

171 . EXTREMAL EIGENVALUES OF RANDOM KERNEL MATRICES WITH POLYNOMIAL SCALING

[02.E1.I41, (page 23)]

Sagnik NANDY, *University of Chicago*
David KOGAN, *Yale University*
Jiaoyang HUANG, *University of Pennsylvania*

We study the spectral norm of random kernel matrices with polynomial scaling, where the number of samples scales polynomially with the data dimension. In this regime, Lu and Yau (2022) proved that the empirical spectral distribution converges to the additive free convolution of a semicircle law and a Marcenko-Pastur law. We demonstrate that under certain assumptions on the distribution of the features, the random kernel matrix can be decomposed into a "bulk" part and a low-rank part. The spectral norm of the "bulk" part almost surely converges to the edge of the limiting spectrum.

In the special case where the random kernel matrices correspond to the inner products of random tensors, the empirical spectral distribution converges to the Marcenko-Pastur law. We prove that the largest and smallest eigenvalues converge to the corresponding spectral edges of the Marcenko-Pastur law.

172. Concentration inequalities for correlated network-valued processes with applications to community estimation and changepoint analysis

[02.M1.C1, (page 16)]

Anirban NATH, *Columbia University*

Network-valued time series are currently a common form of network data. However, the study of the aggregate behavior of network sequences generated from network-valued stochastic processes is relatively rare. Most of the existing research focuses on the simple setup where the networks are independent (or conditionally independent) across time, and all edges are updated synchronously at each time step. In this paper, we study the concentration properties of the aggregated adjacency matrix and the corresponding Laplacian matrix associated with network sequences generated from lazy network-valued stochastic processes, where edges update asynchronously, and each edge follows a lazy stochastic process for its updates independent of the other edges. We demonstrate the usefulness of these concentration results in proving consistency of standard estimators in community estimation and changepoint estimation problems. We also conduct a simulation study to demonstrate the effect of the laziness parameter, which controls the extent of temporal correlation, on the accuracy of community and changepoint estimation.

173 . Who Is Winning? Determining Whether a Candidate Leads in a Ranked-Choice Election

[Special Invited Session 4, (page 28)]

Dan NETTLETON, *Iowa State University*
Shigeki KANAMORI, *Iowa State University*

In an election with more than two candidates, it can be surprisingly complicated to determine whether a candidate is leading from the results of a survey. Even under the simplifying assumption of a simple random sample from an effectively infinite population, there are interesting statistical aspects to consider. Testing whether a particular candidate leads in the population involves a likelihood ratio test whose asymptotic null distribution is a chi-square mixture of the type arising in order-restricted inference. Complexity increases for elections in which each voter is asked to rank candidates on a ballot rather than simply choosing a most preferred candidate. In such ranked-choice elections, instant-runoff voting is often used to determine a winner. We explore how to test whether a particular candidate leads in the voting population based on candidate rankings provided by a sample of voters. We discuss a likelihood ratio test, an intersection-union test, and a simple Bayesian approach for evaluating whether a candidate leads a multi-candidate race in an election that uses ranked-choice and instant-runoff voting.

174 . Assumption-lean weak limit of weighted IPW estimator for two-stage adaptive experiments

[Student Paper Competition 1, (page 5)]

Ziang NIU, *University of Pennsylvania*
Zhimei REN, *University of Pennsylvania*

Adaptive experiments are increasingly indispensable across a wide range of real-world applications, prized for their capacity to minimize welfare loss and enhance practical efficiency. Yet despite their growing use, the statistical foundations for valid inference in such adaptive settings remain underdeveloped. To address this critical gap, we establish a set of new weak convergence results for the widely adopted weighted inverse probability weighting (IPW) estimator when data is collected from two-stage adaptive experiments. Our results stand in stark contrast to prior work: they require significantly weaker assumptions and sharply delineate the phase transitions in weak limits across signal regimes, driven by the adaptive data collection process. Crucially, our framework bridges previously fragmented literatures, unifying them through the lens of signal strength. Building on this theoretical foundation, we

introduce a computationally efficient, provably valid plug-in bootstrap method to derive critical values under the null hypothesis in a statistical testing setup. Extensive simulations and semi-synthetic studies underscore the practical impact of our findings, revealing several statistical behaviors unique to adaptive experiments that are absent in classical randomized designs.

175. A Pseudo-likelihood Approach to Under-5 Mortality Estimation

[03.M1.I48, (page 26)]

Taylor OKONEK, *Macalester College*
Katherine WILSON, *University of Washington*
Jon WAKEFIELD, *University of Washington*

Accurate and precise estimates of the under-5 mortality rate (U5MR) are an important health summary for countries. However, full survival curves allow us to better understand the pattern of mortality in children under five. Modern demographic methods for estimating a full mortality schedule for children have been developed for countries with good vital registration and reliable census data, but perform poorly in many low- and middle-income countries (LMICs). In these countries, the need to utilize nationally representative surveys to estimate the U5MR requires additional care to mitigate potential biases in survey data, acknowledge the survey design, and handle the usual characteristics of survival data, for example, censoring and truncation. We develop parametric and non-parametric pseudo-likelihood approaches to estimating child mortality across calendar time from complex survey data. We show that the parametric approach is particularly useful in scenarios where data are sparse and parsimonious models allow efficient estimation. We compare a variety of parametric models to two existing methods for obtaining a full survival curve for children under the age of 5, and argue that a parametric pseudo-likelihood approach is advantageous in LMICs. We apply our proposed approaches to survey data from four LMICs.

176. Goodness of Fit Testing with Saddlepoint Approximation for Degradation Data

[03.A1.I60, (page 32)]

Lochana PALAYANGODA, *Assistant Professor at University of Nebraska at Omaha*
Hon Keung Tony NG, *Professor at Bentley University*
Aziz GAFUROV, *MS Student at the University of Nebraska at Omaha*

In degradation data modeling, parametric stochastic processes such as gamma, inverse-Gaussian, and Wiener processes are commonly employed to estimate the first-passage time distribution. Accurately identifying the underlying stochastic process is essential to prevent model misspecification errors. Traditional goodness-of-fit tests, including the Kolmogorov–Smirnov, Cramèr–von Mises, and Anderson-Darling tests, are applicable when degradation measurements are taken at equally spaced time intervals. However, in many practical scenarios, degradation data are collected at unequal time intervals due to experimental constraints or missing measurements. This makes conventional goodness-of-fit testing challenging as degradation differences become non-identically distributed. This study introduces novel goodness-of-fit tests test procedures designed to assess the suitability of stochastic degradation models when data are measured at irregular time intervals. Rather than testing the distribution of the degradation measurements directly, the proposed approach evaluates the first-passage time distribution to derive test statistics. The empirical distribution of the first-passage time is estimated using the empirical saddlepoint approximation method. Monte Carlo simulations and practical data applications are conducted to evaluate the Type-I error rates and statistical power of the proposed tests.

177. Bayesian Optimal Phase II Randomized Clinical Trial Design for Immunotherapy with Delayed Outcomes

[03.M1.I50, (page 27)]

Haitao PAN, *St Jude*

TBA

178. State evolution beyond first order methods

[02.M2.I29, (page 19)]

Ashwin PANANJADY, *Georgia Tech*
Michael CELENTANO, *OpenAI*
Chen CHENG, *Stanford University*
Kabir VERCHAND, *Georgia Tech*

We consider the dynamics of iterative optimization algorithms when applied to instances with high-dimensional, random data. When the algorithm of choice is a first-order method, it is known that the dynamics of the method are well approximated by a low-dimensional deterministic recursion known as state evolution. In this paper, we move beyond first-order methods and develop a rigorous state evolution for a far larger set of algorithms. We show that this state evolution can be written in a "canonical form", allowing us to argue existence and uniqueness of the

deterministic updates. Along the way, we develop a variant of Bolthausen's conditioning method that relies on a sequential variant of Gordon's Gaussian comparison inequality.

179. Inference with Randomized Regression Trees

[01.E1.I15, (page 12)]

Snigdha PANIGRAHI, *University of Michigan*
Soham BAKSHI, *University of Michigan*
Yiling HUANG, *University of Michigan*
Walter DEMPSEY, *University of Michigan*

Regression trees are a widely used machine learning algorithm that fit piecewise constant models by recursively partitioning the predictor space. In this talk, I will introduce Randomized Regression Trees (RRT), a novel method that enables statistical inference in the fitted, non-linear tree model. The RRT method achieves this by adding independent Gaussian noise to the gain function underlying the splitting rules of classical regression trees.

The RRT method utilizes the added randomization to obtain an exact pivot using the full dataset, while taking into account the data-dependent structure of the fitted tree. With a small amount of randomization, the RRT method achieves predictive accuracy similar to a tree model trained on the entire dataset. At the same time, it provides significantly more powerful inference than data splitting methods, which rely only on a held-out portion of the data for inference. On various numerical examples, we show how the RRT approach transforms a purely predictive tree algorithm into one capable of delivering statistical inference in the fitted model.

180. Proximal Causal Inference for Conditional Separable Effects

[01.A1.I6, (page 9)]

Chan PARK, *University of Illinois Urbana-Champaign*
Mats STENSRUD, *EPFL*
Eric TCHETGEN TCHETGEN, *University of Pennsylvania*

Scientists regularly pose questions about treatment effects on outcomes conditional on a post-treatment event. However, defining, identifying, and estimating causal effects conditional on post-treatment events requires care, even in perfectly executed randomized experiments. Recently, the conditional separable effect (CSE) was proposed as an interventionist estimand that corresponds to scientifically meaningful questions in these settings. However, while being a single-world estimand, which can be queried experimentally, existing identification results for the CSE require no unmeasured confounding between the outcome and post-treatment event. This assumption can be violated in many applications. In this work, we address this concern by developing new identification and estimation results for the CSE in the presence of unmeasured confounding. We establish nonparametric identification of the CSE in observational and experimental settings when time-varying confounders are present, and certain proxy variables are available for hidden common causes of the post-treatment event and outcome. For inference, we characterize an influence function for the CSE under a semiparametric model in which nuisance functions are a priori unrestricted. Moreover, we develop a consistent, asymptotically linear, and locally semiparametric efficient estimator of the CSE using modern machine learning theory. We illustrate our framework with simulation studies and a real-world cancer therapy trial.

181. TBD

[01.M2.I1, (page 3)]

Gunwoong PARK, *Seoul National University*

TBA

182. Design-based inference for incomplete block designs

[04.M1.I70, (page 36)]

Nicole PASHLEY, *Rutgers University*
Taehyeon KOO,

Design-based inference relies on the random assignment of units into treatment arms as the basis for inference, avoiding standard model-based assumptions. This talk develops novel tools for conducting finite-population design-based inference for complex experiments, focusing on incomplete block designs. These designs are a natural alternative to the complete block design when resource or other constraints limit the number of treatments that can be assigned within a block. To assist practitioners in understanding the trade-offs of using these designs, precision comparisons are made to standard estimators for the complete block, cluster-randomized, and completely randomized designs.

183. Revisiting empirical risk minimization: new risk characterizations and suboptimality results

[02.E1.I40, (page 23)]

Reese PATHAK, *UC Berkeley*

Empirical risk minimization is the optimization workhorse that underlies most of modern statistical

and machine learning practice. However—perhaps somewhat surprisingly—in many situations we still do not fully understand its behavior: for instance, for particular parameter spaces and/or in the presence of covariates. In this talk, we describe two suboptimality results for empirical risk minimization under square loss, in regression settings. We first investigate sparse linear regression when the covariate distribution is anisotropic. In this case, it turns out that there are anisotropic designs for which classical empirical risk minimization methods (e.g. the constrained or penalized Lasso) are rate suboptimal. Second, we study the empirical risk minimizer in the classical Gaussian sequence model over norm-p balls. Here, we show that for p between 1 and 2 there are norm-p bounded signals for which the empirical risk minimizer is minimax rate-suboptimal. We conclude by discussing the consequences of these suboptimality results for the more general problem of learning under a form distribution shift known as covariate shift.

Based on joint work with Cong Ma (U. Chicago), Annie Ulichney (Berkeley), Liviu Aolaritei (Berkeley), and Michael I. Jordan (Berkeley / INRIA).

184 . Uncertainty Quantification for Functionals in Constrained Inverse Problems

[01.E1.I12, (page 11)]

Pratik PATIL, *University of California, Berkeley*

Ill-posed inverse problems arise frequently in modern scientific applications and are characterized by a large set of parameters that are consistent with the observed data. Rigorous non-asymptotic uncertainty quantification in these problems is possible by inferring specific functionals of the unknown parameter under known physical constraints. The presence of the functional, constraints, and potential unidentifiability makes this a challenging inference problem.

This talk will describe a novel framework for constructing confidence intervals in this setting by inverting a specialized (non-standard) likelihood-ratio test. Since critical values for this test typically lack closed-form expressions, we propose calibrating the test on a data-adaptive compact subset of the parameter space, chosen to contain the true parameter with high probability. To make this feasible, we introduce concrete computational strategies based on optimization and custom-designed sampling algorithms. We will first illustrate these methods using simple, low-dimensional examples, followed by a realistic, moderately high-dimensional particle unfolding problem from high-energy physics, demonstrating that our intervals achieve substantially reduced length while maintaining correct finite-sample coverage. Finally, we will discuss promising directions for extending this inference framework to ultra-high-dimensional inverse problems.

185. TBD

[03.E1.I66, (page 35)]

Rohit PATRA, *Linkedin Inc*

TBA

186. Decoding Spatial Tissue Architecture: A Scalable Bayesian Topic Model for Multiplexed Imaging Analysis

[02.A1.I37, (page 22)]

Xiyu PENG, *Texas A&M University*
James SMITHY, *Memorial Sloan Kettering Cancer Center*
Katherine PANAGEAS, *Memorial Sloan Kettering Cancer Center*
Ronglai SHEN, *Memorial Sloan Kettering Cancer Center*

Recent progress in multiplexed tissue imaging is advancing the study of tumor microenvironments to enhance our understanding of treatment response and disease progression. Despite its popularity, there are significant challenges in data analysis, including high computational demands that limit feasibility for large-scale applications and the lack of a principled strategy for integrative analysis across images. To overcome these challenges, we introduce a spatial topic model designed to decode high-level spatial architecture across multiplexed tissue images. Our method integrates both cell type and spatial information within a topic modelling framework, originally developed for natural language processing and adapted for computer vision. We benchmarked its performance through various case studies using different single-cell spatial transcriptomic and proteomic imaging platforms across different tissue types. We show that our method runs significant faster on large-scale image datasets, along with high precision and interpretability. We also demonstrate it consistently identifies biologically and clinically significant spatial "topics", such as tertiary lymphoid structures.

187 . Statistical Guarantees for Semi-Implicit Variational Inference

[03.E1.I65, (page 34)]

Sean PLUMMER, *University of Arkansas*

Transformation based variational families, such

Semi-Implicit Variational Inference (SIVI), show great empirical performance, but their use in applications as an alternative to Markov Chain Monte Carlo is limited due to a lack statistical guarantees. Utilizing the -Variational Bayes framework we delineate conditions under which Gaussian semi-implicit variational families of arbitrary parameter dimension achieve both optimal approximation quality and optimal variational Bayesian risk bounds as well as open directions for future work.

188. Beyond the Odds: Fitting Logistic Regression with Missing Data in Small Samples

[04.M2.I74, (page 38)]

Vivek PRADHAN, *Pfizer Inc., Cambridge, MA 02139, USA*

Logistic regression is widely used across various fields, particularly in the biomedical sciences, where small sample sizes and missing data often present significant challenges. Missing data can severely compromise the validity of statistical analyses, especially in logistic regression, which is sensitive to incomplete covariate information. Statistical science offers several methods to address these issues, with likelihood-based approaches providing a robust solution. These methods modify the likelihood function based on the type of missingness, such as Missing at Random (MAR). Ibrahim (1990) introduced a likelihood-based approach for generalized linear models with missing categorical covariates under MAR. Building on this foundation, Ibrahim and his colleagues developed a series of methodologies for fitting models with missing values, as summarized in the review papers by Ibrahim (2005) and Horton (1999). These works primarily address missing covariates, but fitting logistic regression models with small sample sizes often requires additional steps, such as bias correction, to improve estimation accuracy. This article summarizes Ibrahim (1990)'s method for handling missing categorical covariates and its associated bias correction approach, which helps mitigate bias in small-sample settings. Additionally, it discusses Ibrahim (1996)'s method for handling missing responses, along with the corresponding bias correction. The performance of these methods is evaluated through simulation studies, comparing their effectiveness in addressing bias and separation issues, particularly in small samples. Finally, we provide an R package, **glmFitMiss**, to make these methodologies accessible for practical applications.

189 . Bayesian Modeling of Misclassification Matrices forImproved Verbal Autopsy-Based Mortality Estimatesin LMICs

[02.M1.I25, (page 17)]

Sandipan PRAMANIK, *Johns Hopkins Bloomberg School of Public Health*
Scott ZEGER, *Johns Hopkins Bloomberg School of Public Health*
Dianna BLAU, *Global Health Center, US Centers for Disease Control and Prevention*
Abhirup DATTA, *Johns Hopkins Bloomberg School of Public Health*

Verbal autopsy (VA) algorithms are widely used in low- and middle-income countries (LMICs) to determine individual causes of death (COD), which are then aggregated to estimate population-level mortality crucial for public health policymaking. However, VA algorithms often misclassify COD, leading to biased mortality estimates. A recent method, VA-calibration, aims to correct this bias by incorporating the VA misclassification rate derived from limited labeled COD data collected in the CHAMPS project. Due to limited labeled samples, data are pooled across countries to enhance estimation precision, implicitly assuming uniform misclassification rates. In this presentation, I will highlight substantial cross-country heterogeneity in VA misclassification. This challenges the homogeneity assumption and increases bias. To address this issue, I propose a comprehensive framework for modeling country-specific misclassification matrices in data-scarce settings. The framework introduces an innovative base model that parsimoniously characterizes the misclassification matrix using two latent mechanisms: intrinsic accuracy and systematic preference. We establish that these mechanisms are theoretically identifiable from the data and manifest as invariance in misclassification odds, a pattern observed in CHAMPS data. Building on this foundation, the framework integrates cross-country heterogeneity through interpretable effect sizes and employs shrinkage priors to balance the bias-variance tradeoff in misclassification matrix estimation. This enhances the applicability of VA-calibration and strengthens ongoing efforts to leverage VA for mortality surveillance. I will illustrate these advancements through applications to projects such as COMSA in Mozambique and CA CODE.

190. Examining Directional Association between Depression and Anxiety

[02.E1.I39, (page 23)]

Soumik PURKAYASTHA, *University of Pittsburgh*
Peter X.-K. SONG, *University of Michigan*

Depression and anxiety are debilitating and prevalent diagnoses with wide-reaching negative psychological and economic impacts. Clinicians note that depression and anxiety, although distinct conditions, often occur together in patients, with little information explaining such comorbidity. In absence of information on the underlying aetiology of these diseases, some clinicians hypothesize that one trait may predispose another, thereby inducing a direction of dependence between these psychological traits.

The Intern Health Study (IHS) examines self-reported depression and anxiety among doctors in residency programs in the US. Being able to establish a sense of directionality between anxiety and depression to understand the dominance between these two mental health outcomes is critical to develop adequate clinical diagnostics and administer medical intervention. We propose a novel information-theoretic coefficient that leverages Shannon's entropy metric used to examine directed dependence between anxiety and depression. The proposed method is evaluated by simulation studies and applied to IHS data, where a dominating effect of depression on anxiety is observed in medical interns.

191. Advancing Drug Development with Statistical Innovation - My Perspective

[Special Invited Session 1, (page 9)]
Yongming QU, *Eli Lilly and Company*

The drug development process is both time-consuming and costly, with phase 2 and phase 3 trials often posing significant financial and operational challenges. Recent research has shown that improving the technical success rates in these phases is key to reducing overall development costs. This presentation highlights several initiatives aimed at driving statistical innovation to enhance decision-making and improve efficiency across drug development phases. A few examples will be provided: • Adjustment for the selection bias in early phase studies to improve the decisionmaking • Developing a novel prediction model designed to accelerate early-phase development, in the areas of diabetes and obesity • A case study on tirzepatide, where an efficient dose estimation method was applied to improve dose selection for phase 3 development • Advances in estimands, missing data imputation, and enhanced data collection strategies to ensure the integrity of clinical trials

192. Modeling Spatial Extremes using Non-Gaussian Spatial Autoregressive Models via Convolutional Neural Networks

[Student Paper Competition 2, (page 6)]
Sweta RAI, *Colorado School of Mines*
Sweta RAI, *Colorado School of Mines*
Douglas NYCHKA, *Colorado School of Mines*
Soutir BANDYOPADHYAY, *Colorado School of Mines*

Data derived from remote sensing or numerical simulations often have a regular gridded structure and are large in volume. However, it is challenging to find accurate spatial models that can fill in missing grid cells or simulate the process effectively, especially when there is spatial heterogeneity and heavy-tailed marginal distributions. One effective method is to use a spatial autoregressive (SAR) model, which maps a location and its neighbors to spatially independent random variables. This model is flexible and well-suited for non-Gaussian fields, providing simpler interpretability. In this study, we incorporate the SAR model with Generalized Extreme Value (GEV) distribution innovations –a heavy-tailed distribution– and introduce nonlinear maps that combine a central grid location with its neighbors, capturing extreme spatial behavior based on the heavy-tailed innovations. While these models are fast to simulate due to the sparseness of the construction, the estimation process is slow because the likelihood is intractable. To overcome this, we suggest training a convolutional neural network (CNN), optimized for pattern recognition, on a large training set that spans a useful parameter space, then using the trained network for fast estimation. We apply this model to analyze annual maximum precipitation derived from ERA-Interim-driven WRF simulations within the NA-CORDEX project, enabling us to explore spatial extremal behavior across North America.

193. Proximal Causal Inference with Some Invalid Proxies

[01.E1.I14, (page 12)]
Prabrisha RAKSHIT, *University of Pennsylvania*
Eric TCHETGEN TCHETGEN, *University of Pennsylvania*
Xu SHI, *University of Michigan*

Proximal causal inference is a recently proposed framework to identify and estimate the causal effect of an exposure on a given outcome, in the presence of hidden confounders for which proxies are available. Specifically, proximal inference relies on having observed two valid types of proxies; a *treatment-confounding proxy* related to the outcome only to the

extent that it is associated with an unmeasured confounder, and an *outcome-confounding proxy* related to the treatment only through its association with an unmeasured confounder. Therefore, valid proxies must satisfy stringent exclusion restrictions; mainly, a treatment-confounding proxy must not cause the outcome, while an outcome-confounding proxy must not be caused by the treatment. In order to improve the prospects for identification and possibly the efficiency of the approach, multiple proxies will often be used, raising concerns about bias due to a possible violation of the required exclusion restrictions. To address this concern, we introduce necessary and sufficient conditions for identifying causal effects in the presence of many confounding proxies, some of which may be invalid. Specifically, under a canonical proximal linear structural equations model, we propose a LASSO-based median estimator of the causal effect of primary interest, which simultaneously selects valid proxies and estimates the causal effect with corresponding theoretical performance guarantees. Despite its strengths, the LASSO-based approach can under certain conditions lead to inconsistent treatment proxy selection. To overcome this limitation, we introduce an adaptive LASSO-based proximal estimator, which incorporates adaptive weights to differentially penalize separate treatment proxy coefficients with respect to the ℓ_1 penalty. We formally establish that the adaptive estimator is $\sqrt{n}$-consistent for the causal effect, and when a valid outcome-confounding proxy is available, we construct corresponding asymptotically valid confidence intervals for the causal effect. All theoretical results are supported by extensive simulation studies. We apply the proposed methods to assess the impact of right heart catheterization (RHC) on 30-day survival outcomes for critically ill ICU patients, utilizing data from the SUPPORT study.

194. Box-Cox Transformation - 60 Years After

[01.A1.I9, (page 10)]

Marepalli RAO, *University of Cincinnati*
Nisha SHESHASHAYEE, *University of Cincinnati*
Wedad ALATEBI, *University of Cincinnati*
Tianyuan GUAN, *Kent State University*

The talk will begin with a panoramic introduction to the field of TRANSFORMATIONS starting with the work of M S Bartlett and passing along the way with the work of J W Tukey culminating with a seminal work of David Cox and George Box, who introduced their stylistic transformation proming the moon. Hundreds papers have been published and continue to be published extolling Box-Cox transformations since then. It is time to take stock of their work, sixty years later, critiquing the promises they made.

195. Large Deviation Results for Time Averages of a Rapidly Evolving Dynamic W-Random Graph Model

[01.E1.I16, (page 13)]

Souvik RAY, *School of Data Science & Society, University of North Carolina at Chapel Hill*
Amarjit BUDHIRAJA, *University of North Carolina at Chapel Hill*
Shankar BHAMIDI, *University of North Carolina at Chapel Hill*

Large deviation results for random graph models have been a topic of huge interest following the proof the Large Deviation Principle (LDP) for Erdos-Renyi random graphs by Chatterjee et al.(2010), which was later extended to W-random graphs by Dhara et al.(2022). A dynamic model for W-random graphs can be defined as follows : We attach independent Poisson clocks to each edge of the complete directed graph and then switch on/off these edges according to these clocks. An LDP for the sample path of this dynamic model was recently proved by Braunsteins et al.(2023). In this talk, we shall investigate an LDP for the time-average of this dynamic model when the edge connections are changing very rapidly and this will give rise to a very different looking rate function. The proof of this LDP will employ weak convergence methods which also allows the dynamics to be inhomogeneous with respect to time and space. We shall also discuss an LDP for an interacting particle system defined on this evolving graph model.

196. Leveraging External Data for Testing Heterogenous Treatment Effects in Randomized Clinical Trials

[03.A1.I57, (page 31)]

Boyu REN, *McLean Hospital*
Sandra FORTINI, *Department of Decision Sciences, Bocconi University*
Ventz STEFFEN, *Division of Biostatistics, University of Minnesota*
Lorenzo TRIPPA, *Department of Biostatistics, Harvard T.H. Chan School of Public Health*

In oncology the efficacy of novel therapeutics often differs across patient subgroups, and these variations are difficult to predict during the initial phases of the drug development process. The relation between the power of randomized clinical trials (RCTs) and heterogeneous treatment effects (HTEs) has been

discussed by several authors. In particular, false negative results are likely to occur when the treatment effects concentrate in a subpopulation but the study design did not account for potential HTEs. The use of external data (ED) from completed clinical studies and electronic health records has the potential to improve decision-making throughout the development of new therapeutics, from early-stage trials to registration. Here we discuss the use of ED to evaluate experimental treatments with potential HTEs. We introduce a permutation procedure to test, at the completion of a RCT, the null hypothesis that the experimental therapy does not improve the primary outcomes in any subpopulation. The permutation test leverages the available ED to increase power. Also, the procedure controls the false positive rate at the desired α-level without restrictive assumptions on the ED, for example, in scenarios with unmeasured confounders, different pre-treatment patient profiles in the RCT population compared to the ED, and other discrepancies between the trial and the ED. We illustrate that the permutation test is optimal according to an interpretable criteria and discuss examples based on asymptotic results and simulations, followed by a retrospective analysis of individual patient-level data from a collection of glioblastoma clinical trials.

197. Adaptive Experimental Design Using Shrinkage Estimators

[01.E1.I19, (page 14)]

Evan ROSENMAN, *Claremont McKenna College*
Kristen HUNTER, *University of New South Wales*

In the setting of multi-armed trials, adaptive designs are a popular way to increase estimation efficiency, identify optimal treatments, or maximize rewards to individuals. Recent work has considered the case of estimating the effects of K active treatments, relative to a control arm, in a sequential trial. Several papers have proposed online versions of the classical Neyman allocation scheme to assign treatments as individuals arrive, typically with the goal of using Horvitz-Thompson-style estimators to obtain causal estimates at the end of the trial. However, this approach may be inefficient in that it fails to borrow information across the treatment arms.

In this paper, we consider adaptivity when the final causal estimation is obtained using a Stein-like shrinkage estimator for heteroscedastic data. Such an estimator shares information across treatment effect estimates, providing provable reductions in expected squared error loss relative to estimating the causal effects in isolation. Moreover, we show that the expected loss takes the form of a Gaussian quadratic form, allowing it to be computed efficiently using numerical integration. This paves the way for sequential adaptivity, allowing treatments to be assigned to minimize the shrinker loss. Through simulations, we demonstrate that this can yield meaningful reductions in estimation error. We also characterize how our adaptive algorithm assigns treatments differently than would a sequential Neyman allocation.

198. Preference Optimization on Pareto Sets: On a Theory of Multi-Objective Optimization

[01.E1.I18, (page 13)]

Abhishek ROY, *Texas A&M University*
Geelon SO, *University of California San Diego*
Yi-An MA, *University of California San Diego*

In multi-objective optimization, a single decision vector must balance the trade-offs between many objectives. Solutions achieving an optimal trade-off are said to be Pareto optimal—these are decision vectors for which improving any one objective must come at a cost to another. Many decisions can be Pareto optimal, so when the decision maker can choose only one, this raises questions of which solution to pick and how. We formulate the problem as *Pareto-constrained optimization*, where the goal is to optimize a preference function constrained to a set of Pareto optimal solutions.

This constrained optimization problem poses significant challenges: not only is the constraint set defined implicitly, but it is also generally non-convex and non-smooth, even when the objectives are strongly convex. We propose a reformulation of the problem where the constraint set is redefined in terms of an appropriately defined manifold. This reformulation allows us to introduce a clearer and more accurate notion of optimality and stationarity, improving upon existing definitions in the literature. We provide an algorithm with a last-iterate convergence rate of $O(K^{-1/2})$ to this notion of stationarity when the preference function is Lipschitz smooth and when the objective functions are strongly convex and Lipschitz smooth. Moreover, motivated by practical applications like Reinforcement Learning with Human Feedback (RLHF), we extend this algorithm to tackle the case where the preference function can only be queried via dueling feedback.

199. Bayesian Cooperative Learning for Multimodal Integration

[Student Poster Competition, (page 8)]

Saptarshi ROY, *Texas A&M University*
Sreya SARKAR, *University of Iowa*
Himel MALLICK, *Cornell University*
Nengjun YI, *University of Alabama, Birmingham*

Multimodal integration has made significant strides in recent years, evolving from early to late fusion approaches and achieving notable performance gains over single-view methods. Substantial questions remain, however, particularly at the intersection of dependence-aware multimodal integration and scalable feature selection - both challenging for current integration paradigms. Developing flexible multimodal methods also requires accounting for the grouping structure inherent in multiview data within and across views, a challenge not addressed by published methods. To bridge these longstanding gaps, we propose a flexible Bayesian cooperative learning method, BayesCOOP, which combines Bayesian group spike-and-slab L_1 regularization with intermediate fusion. As one of the first group sparsity-aware multimodal approaches in the field, BayesCOOP significantly outperforms state-of-the-art approaches, including early, late, and intermediate fusion. Analyzing two published multimodal datasets using BayesCOOP, we show that it can be up to 20 times more powerful than existing methods and disclose multimodal discoveries that otherwise cannot be revealed by existing approaches. Our open-source software is publicly available.

200 . Informed MCMC for spatial GLMMs

[03.A1.I58, (page 31)]

Vivekananda ROY, *Iowa State University*

A Riemannian geometric framework for Markov chain Monte Carlo (MCMC) is developed where using the Fisher-Rao metric on the manifold of probability density functions (pdfs) informed proposal densities for Metropolis-Hastings (MH) algorithms are constructed. We exploit the square-root representation of pdfs under which the Fisher-Rao metric boils down to the standard L2 metric, resulting in a straightforward implementation of the proposed geometric MCMC methodology. Unlike the random walk MH that blindly proposes a candidate state using no information about the target, the geometric MH algorithms effectively move an uninformed base density (e.g., a random walk proposal density) towards different global/local approximations of the target density. We will discuss an application of the proposed general geometric framework to construct efficient MCMC algorithms for fitting spatial generalized linear mixed models.

201 . Bayesian randomized basket trial design: a case study from the ultra-rare invasive mould infections

[02.E1.I39, (page 23)]

Satrajit ROYCHOUDHURY, *Pfizer Inc.*

Invasive mould infections (IMIs) are rare but life-threatening. Regulatory approval for new antifungal drugs requires a well-powered, randomized non-inferiority trial, which is nearly infeasible due to the rarity of IMIs. Additionally, heterogeneity among mould types complicates study design and treatment effect interpretation when a study includes patients infected by different types of pathogens. Despite the success of single-arm oncology basket trials in evaluating treatment effect in multiple disease types, statistical methods for randomized basket trials in non-oncologic settings remain underdeveloped. We propose a robust borrowing strategy to enhance the efficiency of randomized basket trials for IMIs by (i) borrowing treatment effects across mould types while accounting for heterogeneity and (ii) augmenting control arms using external data. The proposed approach increases the efficiency and precision of the treatment effect estimates for various moulds. It also increases the ethical appeal by reducing the number of patients required for the control arm. Using simulation and real-life examples, we demonstrated the proposed approach can significantly increase statistical power and precision while maintaining the family-wise type I error rate at an acceptable level. Our approach offers a substantial improvement over the current practice of pooling different moulds together for inference and is applicable to rare disease trials facing similar accrual and ethical challenges.

202. A Poincaré Inequality and Consistency Results for Signal Sampling on Large Graphs

[02.E1.I42, (page 24)]

Luana RUIZ, *Johns Hopkins University*
Thien LE,
Stefanie JEGELKA,

Large-scale graph machine learning is challenging as the complexity of learning models scales with the graph size. Subsampling the graph is a viable alternative, but sampling on graphs is nontrivial as graphs are non-Euclidean. Existing graph sampling techniques require not only computing the spectra of large matrices but also repeating these computations when the graph changes, e.g., grows. In this paper, we introduce a signal sampling theory for a type of graph limit—the graphon. We prove a Poincaré inequality for graphon signals and show that complements

of node subsets satisfying this inequality are unique sampling sets for Paley-Wiener spaces of graphon signals. Exploiting connections with spectral clustering and Gaussian elimination, we prove that such sampling sets are consistent in the sense that unique sampling sets on a convergent graph sequence converge to unique sampling sets on the graphon. We then propose a related graphon signal sampling algorithm for large graphs, and demonstrate its good empirical performance on graph machine learning tasks.

203. Two-Level SLOPE: Balancing Simplicity and Effectiveness in Adaptive Regularization

[03.A1.I61, (page 32)]

Cynthia RUSH, *Columbia University*
Zhiqi BU, *Amazon*
Ruijia WU, *SJTU*
Jason KLUSOWSKI, *Princeton University*

Among techniques for high-dimensional linear regression, Sorted L-One Penalized Estimation (SLOPE) generalizes the LASSO via an adaptive L1 regularization that applies heavier penalties to larger coefficients in the model. To achieve such adaptivity, SLOPE requires the specification of a complex hierarchy of penalties, i.e., a monotone penalty sequence in R^p, in contrast to a single penalty scalar for LASSO. Tuning this sequence when p is large poses a challenge, as brute force search over a grid of values is computationally infeasible. In this talk, we introduce the two-level SLOPE, an important subclass of SLOPE, with only three hyperparameters. We demonstrate both empirically and analytically that two-level SLOPE not only preserves the advantages of general SLOPE — such as improved mean squared error and overcoming the Donoho-Tanner power limit — but also exhibits computational benefits by reducing the penalty hyperparameter space.

204 . Approximation rates of entropic maps in semidiscrete optimal transport

[01.M2.I3, (page 4)]

Ritwik SADHU, *Amazon*
Ziv GOLDFELD, *Cornell University*
Kengo KATO, *Cornell University*

Entropic optimal transport offers a computationally tractable approximation to the classical problem. We study the approximation rate of the entropic optimal transport map (in approaching the Brenier map) when the regularization parameter ε tends to zero in the semidiscrete setting, where the input measure is absolutely continuous while the output is finitely discrete. Previous work shows that the approximation rate is $O(\sqrt{\varepsilon})$ under the L^2-norm with respect to the input measure. In this work, we establish faster, $O(\varepsilon^2)$ rates up to polylogarithmic factors, under the dual Lipschitz norm, which is weaker than the L^2-norm. For the said dual norm, the $O(\varepsilon^2)$ rate is sharp. As a corollary, we derive a central limit theorem for the entropic estimator for the Brenier map in the dual Lipschitz space when the regularization parameter tends to zero as the sample size increases.

205. Random forests for binary geospatial data

[02.E1.I45, (page 25)]

Arkajyoti SAHA, *University of California, Irvine*
Abhirup DATTA, *Johns Hopkins Bloomberg School of Public health*

Existing implementations of random forests for binary data cannot explicitly account for data correlation common in geospatial and time-series settings. For continuous outcomes, recent work has extended random forests (RF) to RF-GLS that incorporate spatial covariance using the generalized least squares (GLS) loss. However, adoption of this idea for binary data is challenging due to the use of the Gini impurity measure in classification trees, which has no known extension to model dependence. We show that for binary data, the GLS loss is also an extension of the Gini impurity measure, as the latter is exactly equivalent to the ordinary least squares (OLS) loss. This justifies using RF-GLS for non-parametric mean function estimation for binary dependent data. We then consider the special case of generalized mixed effects models, the traditional statistical model for binary geospatial data, which models the spatial random effects as a Gaussian process (GP). We propose a novel link-inversion technique that embeds the RF-GLS estimate of the mean function from the first step within the generalized mixed effects model framework, enabling estimation of non-linear covariate effects and offering spatial predictions. We establish consistency of our method, RF-GP, for both mean function and covariate effect estimation. The theory holds for a general class of stationary, absolutely regular dependent processes that includes common choices like Gaussian processes with Matérn or compactly supported covariances and autoregressive processes. We demonstrate that RF-GP outperforms competing methods for estimation and prediction in both simulated and real-world data.

206. Structured Bayesian Variable Selection for Microbiome Compositional Data Using Graph-Guided Shrinkage

[03.E1.I67, (page 35)]

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

We propose a Bayesian regression model for microbiome compositional data that accounts for both sparsity and microbial structure. The method employs a structured horseshoe prior that encourages variable selection while borrowing strength across related taxa through a graph-informed shrinkage. To respect compositional constraints, regression coefficients are modeled under a sum-to-zero condition. Posterior inference is performed via an efficient blocked Gibbs sampler with Langevin updates. Through simulation studies and application to real microbiome data, we demonstrate improved feature selection and predictive performance over existing state of the art algorithms.

207. Probabilistic Classification and Uncertainty Quantification of Sahara Desert Climate Using Feedforward Neural Networks

[03.A1.I58, (page 31)]

Indranil SAHOO, *Virginia Commonwealth University*
Stephen TIVENAN, *Virginia Commonwealth University*
Yanjun QIAN, *Virginia Commonwealth University*

Climate classification plays a vital role in agricultural planning, hydrological studies and climate science. One of the most widely used systems for classifying global climate zones is the Köppen-Trewartha (KT) classification. However, the KT classification is fundamentally deterministic, offering discrete labels to spatial locations without accounting for uncertainties in classification. In this paper, we utilize a continuous metric based on the KT classification to enable probabilistic modeling of climatic zones. We implement a feedforward artificial neural network (ANN) for classification and forecasting, allowing for efficient, uncertainty-aware categorization of the climatic regions, thereby offering a more nuanced understanding of transitional climate zones compared to traditional categorical methods. We apply this method to the Sahara Desert region over the 30-year period of 1960 - 1990, using data at more than 367,000 space-time locations from the first 10 years for model training. We assess the model's short- and long-term forecasting capabilities to evaluate its stability and accuracy over time. We also present visual comparisons between the probabilistic outputs of our model and the traditional KT classification. Our analysis not only highlights the temporal evolution of climatic zones across the region but also identifies areas undergoing significant flux, providing insights into broader trends in desertification.

208. Statistical modeling and prediction of patient recruitment in multicenter clinical trials

[04.M2.I74, (page 38)]

Srijata SAMANTA, *Bristol Myers Squibb*

Prediction of study timeline is a critical aspect of clinical trials. Use of resources and statistical power depend on the number of recruited patients in the trial. This necessitates prediction of the number of patients enrolled in a reasonable time frame and within a limited research budget or that of the time at which target enrollment will be achieved. Since the enrollment rate is never identical in various countries/sites, we propose to consider a multi-center clinical trial with "k" centers (countries or sites) and model the waiting times (ti - ti-1) of the i-th center as exponential variables with mean i depending on i. We consider a hierarchical Bayesian model and use the posterior predictive distributions of the future waiting times of all the centers for predictions related to subject accrual. This hierarchical modelling has the scope of "borrowing strength" from other centers while making inference for the parameter of the model for any center. Handling multi-center modelling can also be used to monitor countries enrolling patients faster than expected and therefore putting the diversity plan at risk. Along with the methodology development we also perform retrospective analysis of published and completed clinical studies.

209. TBD

[01.M2.I1, (page 3)]

Kris SANKARAN, *University of Wisconsin-Madison*

TBA

210. Misspecified Yet Credible: A Generalized Bayes Framework for Uncertainty Quantification in High-Dimensional Bayesian Vector Autoregressive Models

[04.M2.I72, (page 37)]

Partha SARKAR, *Florida State University*
Ray BAI, *University of South Carolina*

Vector autoregressive (VAR) models are widely used to capture linear dependencies in multivariate time series, yet Bayesian inferential theory for high-dimensional VAR remains largely undeveloped. We

propose a generalized Bayes framework that automatically adapts to sparsity and is robust to misspecification of both the error distribution and covariance structure. Under mild regularity conditions, we show that this approach yields reliable uncertainty quantification for the VAR transition matrices in very high dimensions. As a corollary, the same strategy also delivers valid inference for sparse high-dimensional stochastic regressions with serially correlated errors.

211 . Bayesian interpretation of a second-order efficient empirical Bayes confidence interval

[Student Poster Competition, (page 8)]

Aditi SEN, *Department of Mathematics, University of Maryland, College Park*
Masayo Y. HIROSE, *Institute of Mathematics for Industry, Kyushu University*
Partha LAHIRI, *Department of Mathematics, University of Maryland, College Park*

Matching priors providing both frequentist and Bayesian justification in an inferential problem have been of interest in literature. Such priors have been explored in the context of point prediction of mixed effects in hierarchical area-level models. We provide for the first time a dual validation in the case of interval estimation in such a context. In this article, we propose a Bayesian interpretation of a highly efficient empirical Bayes confidence interval for small area means, proposed by Yoshimori and Lahiri (2014) (which we denote as I_i^{YL}). The idea of empirical Bayes confidence interval was first proposed by Cox, which was later improved by Yoshimori and Lahiri (2014), in terms of coverage error reduction from $O(m^{-1})$ to $o(m^{-1})$, where m is the number of small areas. The authors carefully devised an adjusted maximum likelihood estimator and based their prediction interval on a frequentist approach. Our aim is to investigate this from the Bayesian viewpoint. To achieve our goal, we consider the variance component (i.e., the hyperparameter) to have a prior distribution and obtain a higher-order asymptotic expansion of the posterior coverage of I_i^{YL}. In the posterior coverage expansion, we show that the prior appears in terms of order $O\_p(m^{-1})$. It is noteworthy that the number of small areas to be pooled may not be very large in practice and hence the effect of this on the prior choice could be substantial. As a result, we derive a prior for which this order $O\_p(m^{-1})$ term vanishes, thereby making the posterior coverage close to the desired empirical Bayes nominal level $(1-\alpha)$. This matching prior, provides a valid Bayesian interpretation of the second-order efficient Cox-type empirical Bayes confidence interval I_i^{YL}. We also show that the matching prior yields a proper posterior under mild regularity conditions. Our theoretical justifications are then corroborated through a Monte Carlo simulation study.

212 . Optimal Use of Survey Weights for Causal Inference under Informative Sampling

[Student Poster Competition, (page 8)]

Shubhajit SEN, *North Carolina State University*
Shu YANG, *Professor of Statistics*

The increasing availability of survey data for causal inference on treatment effects has brought new opportunities, yet most existing methodologies assume ignorability of treatment assignment and non-informative sampling. In practice, survey data often include survey weights, but the sampling mechanism is frequently informative—i.e., not independent of the outcome of interest, conditional on observed covariates and the treatment variable—especially when the detailed design information is undisclosed. Addressing the optimal use of survey weights for causal inference under informative sampling thus remains an open problem. In this work, we demonstrate how survey weights can be effectively leveraged to enhance the efficiency of classic Horvitz-Thompson estimators. Specifically, we derive the efficient influence function within the class of Regular and Asymptotically Linear estimators and propose a novel estimator based on it. Employing a super-population framework, we establish the estimator's doubly robust properties and, using M-estimation theory, prove its asymptotic $\sqrt{N}$-normality under parametric nuisance modeling. To accommodate data-adaptive and flexible machine learning methods, we further extend the theory to demonstrate "rate-doubly robust" convergence. This ensures that the estimator's convergence rate surpasses $1/\sqrt{N}$ when the product of nuisance function convergence rates exceeds $1/\sqrt{N}$. Additionally, we provide a variance estimation framework within a mixed inferential context, accounting for dual sources of uncertainty: the finite population sampled from the super-population and the survey sample drawn from the finite population. Through extensive simulations and an analysis of data from the Medical Expenditure Panel Survey, we illustrate that our proposed estimator consistently outperforms traditional survey-weighted methods in terms of efficiency.

Keywords: Complex survey, Data-Adaptive

Method, Doubly Robust Estimation, Empirical process, Population average treatment effect.

213. Spectral algorithms for community detection in multiview networks

[02.E1.I42, (page 24)]

Subhabrata SEN, *Harvard University*
Xiaodong YANG, *Harvard University*
Yue M. LU, *Harvard University*

In this talk, we will discuss novel spectral algorithms for community detection in multiview networks. We will show that these algorithms attain the sharp thresholds for weak recovery in this problem. Based on joint work with Xiaodong Yang and Yue M. Lu (Harvard).

214 . Bayesian Dynamic Borrowing Meets Machine Learning: A Case Study of Hybrid Control Arms in Oncology Trial

[01.A1.I5, (page 9)]

Sanhita SENGUPTA, *Bristol Myers Squibb*
Jixian WANG, *Bristol Myers Squibb*
Ram TIWARI, *HopeAI*

This study demonstrates how the integration of Bayesian dynamic borrowing and machine learning can transform clinical trial design through the intelligent incorporation of real-world data into randomized controlled trials. Using the CheckMate-057 trial in advanced lung cancer as a retrospective case study, we developed a novel framework that combines Bayesian approaches with propensity score methods to create robust hybrid control arms. Our method reduced trial duration by more than three months while maintaining statistical integrity, achieving hazard ratios consistent with the original trial (HR 0.73). We present a practical blueprint for implementing these methods and address key challenges in data integration and model validation. This work demonstrates how modern statistical approaches can accelerate drug development without compromising scientific rigor—a critical advancement particularly relevant for rare disease and oncology trials where traditional designs face mounting challenges.

215. AI-powered Flexible-design Simulation Assistant for Clinical Trials

[02.M2.I27, (page 18)]

Subhajit SENGUPTA, *Cytel Inc.*
Kyle WATHEN, *Cytel Inc.*
Gabriel POTVIN, *Cytel Inc.*
Anoop Singh RAWAT, *Cytel Inc.*

In this talk, we present AIFS-ACT: AI-powered Flexible-design Simulation Assistant for Clinical Trials. Simulation is a powerful strategy for exploring the space of design parameters, enabling the optimization of clinical trial designs. Commercial platforms like East Horizon from Cytel Inc. are widely adopted for clinical trial design and optimization. However, these tools often lack the flexibility required to accommodate complex or novel trial designs. To address these limitations, we introduced an open-source R integration initiative within our software stack, empowering users to incorporate customized R code into our commercial software. This integration enhanced or replaced native functionalities, enabling more flexible and complex trial simulations.

AIFS-ACT is designed to assist users, especially those who are new to our clinical trial simulation platform — in developing compatible R functions for flexible trial designs. By leveraging generative AI technology, our assistant ensures the seamless creation of R functions that adhere to the required input/output specifications, thereby reducing the learning curve and minimizing errors.

Current key features of AIFS-ACT include: -Simulating Patient Responses: Binary, Continuous, Time-to-event, and Repeated-measure endpoints. -Analyzing Simulated Data: Statistical analysis for the aforementioned endpoints. -Randomization: Flexible randomization of patients across treatment groups. -Enrollment and Dropout Modeling: Custom mechanisms for realistic patient enrollment and dropout scenarios. -Treatment Selection: Supporting Multi-Arm Multi-Stage (MAMS) trial designs.

Our platform integrates Azure OpenAI's GPT-4o, a large language model (LLM), with Cytel's internally developed R package, CyneRgy, to enable the creation of custom adaptive clinical trial designs. AIFS-ACT also includes testing functionality for AI-generated R code, which detects and flags errors to improve robustness. All data interactions adhere to Azure OpenAI's stringent data protection policies to ensure security and compliance.

AIFS-ACT is currently offered to East Horizon users, providing a transformative AI-enhanced tool for the flexible design and simulation of clinical trials, accelerating innovation and improving trial efficiency

Keywords: Generative AI, Large Language Models (LLM), Coding Assistant, R Integration, Adaptive Clinical Trial Design

216 . Bayesian Semi-supervised Inference via a Debiased Modeling Approach

[Student Paper Competition 1, (page 5)]

Gözde SERT, *Texas A&M University*
Gözde SERT, *Texas A&M University*
Abhishek CHAKRABORTTY, *Texas A&M University*
Anirban BHATTACHARYA , *Texas A&M University*

Inference in semi-supervised (SS) settings has received substantial attention in recent years due to increased relevance in modern big-data problems. In a typical SS setting, there is a much larger sized unlabeled data, containing observations only for a set of predictors, in addition to a moderately sized labeled data containing observations for both an outcome and the set of predictors. Such data arises naturally from settings where the outcome, unlike the predictors, is costly or difficult to obtain. One of the primary statistical objectives in SS settings is to explore whether parameter estimation can be improved by exploiting the unlabeled data. A novel Bayesian approach to SS inference for the population mean estimation problem is proposed. The proposed approach provides improved and optimal estimators both in terms of estimation efficiency as well as inference. The method itself has several interesting artifacts. The central idea behind the method is to model certain summary statistics of the data in a targeted manner, rather than the entire raw data itself, along with a novel Bayesian notion of debiasing. Specifying appropriate summary statistics crucially relies on a debiased representation of the population mean that incorporates unlabeled data through a flexible nuisance function while also learning its estimation bias. Combined with careful usage of sample splitting, this debiasing approach mitigates the effect of bias due to slow rates or misspecification of the nuisance parameter from the posterior of the final parameter of interest, ensuring its robustness and efficiency. Concrete theoretical results, via Bernstein–von Mises theorems, are established, validating all claims, and are further supported through extensive numerical studies. To our knowledge, this work is possibly the first on Bayesian inference in SS settings, and its central ideas also apply more broadly to other Bayesian semiparametric inference problems.

217. Sequential change detection with simulators

[01.E1.I12, (page 12)]

Shubhanshu SHEKHAR, *University of Michigan, Ann Arbor*

In this talk, I will describe a simple strategy for sequential change detection that leverages additional information from a simulator or predictive model. Focusing on the problem of detecting changes in the mean of bounded observations, I will recall how this task can be reduced to that of constructing sequential power-one tests. Building on this reduction, I will introduce a new class of sequential tests and detection schemes that efficiently integrate simulator-generated data. I will present some theoretical results that quantify the gains over the no-simulator baseline and then conclude the talk with a discussion of potential directions for future research.

218. Large language models in sports analytics

[03.M1.I47, (page 26)]

Weining SHEN, *University of California, Irvine*

In this talk, I will discuss recent work on evaluating the sports understanding of LLMs, using newly introduced benchmark datasets. Our evaluation covers a range of tasks, from basic queries about rules and historical facts to complex, context-specific reasoning, as well as assessing the sports reasoning capabilities of video language models. Experiments show that models fall short on hard tasks that require deep reasoning and rule-based understanding. We hope the published benchmarks will serve as a critical step toward improving models' capabilities in sports understanding and reasoning.

219 . Compartmentalization of Discrete Repeated Measures in Patient Reported Outcomes (PROs)

[03.A1.I62, (page 32)]

Manasi SHETH, *University of Wisconsin - Whitewater*

There is a recent advancement in the field of statistics to understand the geometry or connectedness of the data due to the massive amounts of data being generated. In biostatistics, or medical field, it is important for patients to have access to high-quality , safe and effective medical devices as well as efficacious treatments. Patient reported outcomes are often relevant in assessing diagnostic evaluations and can be used to capture a patient's everyday experience with a medical device, including experience outside of the clinician's office and the effects of treatment on a patient's activities of daily living and functionality. In some cases, PRO measures enable us to measure important health status information that cannot yet be detected by other measures, such as pain and mobility. Here, I present the geometrical representations of a novel approach for analyzing PROs using Compartmentalization Method.

220 . Sparse Bayesian Group Factor Model for Feature Interactions in Multiple Count Tables Data

[02.M2.I30, (page 19)]
Zhang SHUANGJIE, *Texas A & M University*
Shen YUNING, *Department of Chemical and Biomolecular Engineering, UCLA*
Chen IRENE A., *Department of Chemical and Biomolecular Engineering, UCLA*
Lee JUHEE, *Department of Statistics, UCSC*

Group factor models have been developed to infer relationships between multiple co-occurring multivariate continuous responses. Motivated by complex count data from multi-domain microbiome studies using next-generation sequencing, we develop a sparse Bayesian group factor model (Sp-BGFM) for multiple count table data that captures the interaction between microorganisms in different domains. Sp-BGFM uses a rounded kernel mixture model using a DP prior with log-normal mixture kernels for count vectors. A group factor model is used to model the covariance matrix of the mixing kernel that describes microorganism interaction. We construct a Dirichlet-Horseshoe (Dir-HS) shrinkage prior and use it as a joint prior for factor loading vectors. Joint sparsity induced by a Dir-HS prior greatly improves the performance in high-dimensional applications. We further model the effects of covariates on microbial abundances using regression. The semiparametric model flexibly accommodates large variability in observed counts and excess zero counts.

221. Covariate adjustment for marginal estimates in randomized trials: a primer and extension to interval-censored outcomes.

[02.M1.I23, (page 16)]
Richard SIZELOVE, *Eli Lilly & Company*

We discuss the covariate-adjusted log-rank test, a recent methodological development that offers guaranteed efficiency gains over the standard log-rank test by incorporating baseline covariates while preserving valid inference under randomization. We illustrate the method's flexibility as a general framework through extension to interval-censored data, where failure times are not observed exactly but are known only to lie within intervals.

222 . Correcting Latent Class Confounder Bias in Observational Studies

[02.M1.I23, (page 16)]
Abdul-Nasah SOALE, *Case Western Reserve University*

Model misspecification is a common problem in estimating treatment effects in observational studies. This paper presents the problem of estimating treatment effects in regression involving latent clusters induced by the inclusion of proxy and ill-defined categorical predictors in the model. A model-based subgroup analysis is proposed for recovering the latent classes and correcting the biased in the treatment effect estimation. The proposed method also provides a check for endogeneity between the latent class variable and the observed predictors via sufficient summary plots. The performance of the method on synthetic and real data are included. The theoretical justifications are also provided.

223. Universality of Max-Margin Classifiers

[01.E1.I17, (page 13)]
Youngtak SOHN, *Brown University*
Andrea MONTANARI, *Stanford University*
Feng RUAN, *Northwestern University*
Basil SAEED, *Stanford University*

Maximum margin binary classification is one of the most fundamental algorithms in machine learning, yet the role of featurization maps and the high-dimensional asymptotics of the mis-classification error for non-Gaussian features are still poorly understood. In this talk, we consider the high-dimensional setting where the number of samples and the input dimension is proportional and prove a universality result: the asymptotic misclassification error depends solely on the covariance structure of the feature vectors and their covariance with a latent variable determining class labels. Consequently, the over-parametrization threshold and the mis-classification error can be analyzed using a simpler Gaussian model. The central mathematical difficulty lies in the fact that max-margin is not the maximizer (or minimizer) of an empirical average, but the maximizer of a minimum over the samples.

224. Semi-Parametric Batched Global Multi-Armed Bandits with Covariates

[02.E1.I43, (page 24)]
Hyebin SONG, *Pennsylvania State University*
Sakshi ARYA, *Case Western Reserve University*

The multi-armed bandits (MAB) framework is a widely used approach for sequential decision-making, where a decision-maker selects an arm in each round with the goal of maximizing long-term rewards. Moreover, in many practical applications, such as personalized medicine and recommendation systems, feedback is provided in batches, contextual information is available at the time of decision-making, and

rewards from different arms are related rather than independent. In this talk, I will present a novel semiparametric framework for batched bandits with covariates and a shared parameter across arms, leveraging the single-index regression (SIR) model to capture relationships between arm rewards while balancing interpretability and flexibility. Our algorithm, Batched single-Index Dynamic binning and Successive arm elimination (BIDS), employs a batched successive arm elimination strategy with a dynamic binning mechanism guided by the single-index direction. We consider two settings: one where a pilot direction is available and another where the direction is estimated from data, deriving theoretical regret bounds for both cases. When a pilot direction is available with sufficient accuracy, our approach achieves minimax-optimal rates (with $d = 1$) for nonparametric batched bandits, circumventing the curse of dimensionality. Extensive experiments on simulated and real-world datasets demonstrate the effectiveness of our algorithm compared to the nonparametric batched bandit method.

225. Prediction of Feed Efficiency Traits in Beef Cattle Using Host Genomic and Metagenomic Sequence Data

[03.E1.I63, (page 34)]

Matt SPANGLER, *University of Nebraska-Lincoln*
Matthew SPANGLER, *University of Nebraska-Lincoln*
Andrew LAKAMP, *University of Nebraska-Lincoln*
Samodha FERNANDO, *University of Nebraska-Lincoln*

This study investigated the accuracy of predicting beef cattle feed efficiency related traits using both host genomic and metagenomic features from 717 animals. The phenotypes were average daily dry matter intake and average daily gain. Animal genotypes consisted of 749,922 imputed sequence variants, while metagenomic data comprised 16,583 open reading frames (ORF) from ruminal microbiota. The ORF were analyzed as log-transformed relative abundance. The mixed linear models fitted included either an animal genetic effect only, a metagenomic effect only, both (additive), or both and their interaction. Fixed effects in all models included a management group defined as the concatenation of year-season-diet, linear covariates for breed proportions and expected heterozygosity. The animal genetic effect utilized a genomic relationship matrix while six different metagenomic relationship matrices were explored relative to modelling the metagenomic effect including different methods of scaling (within or across diet types) and by weighting metagenomic features as a function of the estimated heritability of the feature (weight = 1-h2). Two cross-validation schemes were applied to evaluate prediction accuracy: 4-fold cross-validation balanced for diet type and leave-one-diet-out cross-validation. Prediction accuracy was measured as the Pearson correlation between an animal's summed random effects and its phenotype adjusted for fixed effects. Minimal differences in results existed among different formations of the metagenomic relationship matrices. Phenotype prediction accuracy ranged from 0.01 – 0.30. Models which combined genome and metagenome data outperformed those using either data source alone. The rumen metagenome can explain a significant proportion of variation in beef cattle feed efficiency traits. Using both host genomic and metagenomic data jointly leads to more accurate phenotypic predictions. Multiple methods of forming the metagenome (co)variance matrix lead to similar prediction accuracies.

226 . Computational-Statistical Trade-offs in Kernel Two-Sample Testing

[02.E1.I41, (page 23)]

Bharath SRIPERUMBUDUR, *Pennsylvania State University*
Soumya MUKHERJEE, *Pennsylvania State University*

Reproducing Kernel Hilbert Space (RKHS) embedding of probability distributions has proved to be an effective approach, via MMD (maximum mean discrepancy) for nonparametric hypothesis testing problems involving distributions defined over general (non-Euclidean) domains. While a substantial amount of work has been done on this topic, only recently, minimax optimal two-sample tests have been constructed that incorporate, unlike MMD, both the mean element and a regularized version of the covariance operator. However, as with most kernel algorithms, the optimal test scales cubically in the sample size, limiting its applicability. In this work, we propose a spectral-regularized two-sample test based on random Fourier feature (RFF) approximation and investigate the trade-offs between statistical optimality and computational efficiency. We show the proposed test to be minimax optimal if the approximation order of RFF (which depends on the smoothness of the likelihood ratio and the decay rate of the eigenvalues of the integral operator) is sufficiently large. We develop a practically implementable permutation-based version of the proposed test with a data-adaptive strategy for selecting the regularization parameter. Finally, through numerical experiments on simulated and benchmark datasets, we demonstrate that the proposed RFF-based test is computationally efficient

and performs almost similar (with a small drop in power) to the exact test.

227. Mixed Model Trace Regression

[04.M2.I72, (page 37)]

Sanvesh SRIVASTAVA, *The University of Iowa*
Ian HULTMAN, *The University of Iowa*

We introduce mixed model trace regression (MMTR), a mixed model linear regression extension for scalar responses and high-dimensional matrix-valued covariates. MMTR's fixed effects component is equivalent to trace regression, with an element-wise lasso penalty imposed on the regression coefficients matrix to facilitate the estimation of a sparse mean parameter. MMTR's key innovation lies in modeling the covariance structure of matrix-variate random effects as a Kronecker product of low-rank row and column covariance matrices, enabling sparse estimation of the covariance parameter through low-rank constraints. We establish identifiability conditions for the estimation of row and column covariance matrices and use them for rank selection by applying group lasso regularization on the columns of their respective Cholesky factors. We develop an Expectation-Maximization (EM) algorithm extension for numerically stable parameter estimation in high-dimensional applications. MMTR achieves estimation accuracy comparable to leading regularized quasi-likelihood competitors across diverse simulation studies and attains the lowest mean square prediction error compared to its competitors on a publicly available image dataset.

228. Consistency of empirical distributions of sequences of graph statistics in networks with dependent edges

[03.M2.I54, (page 29)]

Jonathan STEWART, *Florida State University*

One of the first steps in applications of statistical network analysis is frequently to produce summary charts of important features of the network. Many of these features take the form of sequences of graph statistics counting the number of realized events in the network, examples of which include the degree distribution, as well as the edgewise shared partner distribution, and more. We provide conditions under which the empirical distributions of sequences of graph statistics are consistent in the ℓ_∞-norm in settings where edges in the network are dependent. We accomplish this task by deriving concentration inequalities that bound probabilities of deviations of graph statistics from the expected value under weak dependence conditions. We apply our concentration inequalities to empirical distributions of sequences of graph statistics and derive non-asymptotic bounds on the ℓ_∞-error which hold with high probability. Our non-asymptotic results are then extended to demonstrate uniform convergence almost surely in selected examples. We illustrate theoretical results through examples, simulation studies, and an application.

229. Quantifying the effects of transfer learning in min-norm interpolation

[02.E1.I44, (page 24)]

Pragya SUR, *Harvard University*
Yanke SONG, *Harvard University*
Sohom BHATTACHARYA, *University of Florida*

Min-norm interpolators naturally emerge as implicit regularized limits of modern machine learning algorithms. Recently, their out-of-distribution risk was studied when test samples are unavailable during training. However, in many applications, a limited amount of test data is typically available during training. The properties of min-norm interpolation in this setting are not well understood. In this talk, I will present a characterization of the generalization error of pooled min-L2-norm interpolation under covariate and model shifts. I will demonstrate that the pooled interpolator captures both early fusion and a form of intermediate fusion. Our results have several implications. Under model shift, adding data always hurts prediction when the signal-to-noise ratio is low. However, for higher signal-to-noise ratios, transfer learning helps as long as the shift-to-signal ratio lies below a threshold that I will define. Our results further show that under covariate shift, if the source sample size is small relative to the dimension, heterogeneity between domains improves the risk. Time permitting, I will introduce a novel anisotropic local law that helps achieve some of these characterizations and may be of independent interest in random matrix theory. This is joint work with Yanke Song and Sohom Bhattacharya.

230. Approximating Distributions via Deep Generative Models: Theory, Limitations and Directions

[03.E1.I65, (page 34)]

Edric TAM, *Stanford University*
Edric TAM, *Stanford University*
David DUNSON, *Duke University*

Deep generative models are routinely used in generating samples from complex, highdimensional distributions. Despite their apparent successes, their statistical properties are not well understood. A com-

mon assumption is that with enough training data and sufficiently large neural networks, deep generative model samples will have arbitrarily small errors in sampling from any continuous target distribution. We set up a unifying framework that debunks this belief. We demonstrate that broad classes of deep generative models, including variational autoencoders and generative adversarial networks, are not universal generators. Under the predominant case of Gaussian latent variables, these models can only generate concentrated samples that exhibit light tails. Using tools from concentration of measure and convex geometry, we give analogous results for more general log-concave and strongly log-concave latent variable distributions. We extend our results to diffusion models via a reduction argument. We use the Gromov–Levy inequality to give similar guarantees when the latent variables lie on manifolds with positive Ricci curvature. These results shed light on the limited capacity of common deep generative models to handle heavy tails. We illustrate the empirical relevance of our work with simulations and financial data.

231. Inference and Learning for Signed Networks Guided by Social Theory

[03.M2.I56, (page 30)]

Weijing TANG, *Carnegie Mellon University*

In many real-world networks, relationships often go beyond simple presence or absence; they can be positive (e.g., friendship, alliance, and mutualism) or negative (e.g., enmity, disputes, and competition). These negative relationships display substantially different properties from positive ones, and more importantly, their presence interacts in unique ways. The balance theory originating from social psychology, illustrated by proverbs like "a friend of my friend is my friend" and "an enemy of my enemy is my friend", provides insight into the formation mechanism of positive and negative connections. In this talk, we characterize the balance theory with a novel and natural notion of population-level balance. We propose a nonparametric inference method to evaluate the real-world evidence of population-level balance in signed networks. Inspired by the empirical findings, we further develop a general latent space framework for modeling signed networks while accommodating the balance theory.

232. Belted and Ensembled Neural Network for Linear and Nonlinear Sufficient Dimension Reduction

[03.E1.I68, (page 35)]

Yin TANG, *Pennsylvania State University*
Bing LI, *Pennsylvania State University*

We introduce a unified, flexible, and easy-to-implement framework of sufficient dimension reduction that can accommodate both linear and nonlinear dimension reduction, and both the conditional distribution and the conditional mean as the targets of estimation. This unified framework is achieved by a specially structured neural network — the Belted and Ensembled Neural Network (BENN) — that consists of a narrow latent layer, which we call the belt, and a family of transformations of the response, which we call the ensemble. By strategically placing the belt at different layers of the neural network, we can achieve linear or nonlinear sufficient dimension reduction, and by choosing the appropriate transformation families, we can achieve dimension reduction for the conditional distribution or the conditional mean. Moreover, thanks to the advantage of the neural network, the method is very fast to compute, overcoming a computation bottleneck of the traditional sufficient dimension reduction estimators, which involves the inversion of a matrix of dimension either p or n. We develop the algorithm and convergence rate of our method, compare it with existing sufficient dimension reduction methods, and apply it to two data examples.

233. Variable Selection in Spatial Regression: Local LASSO

[03.M1.I46, (page 26)]

Debjoy THAKUR, *Washington University in St. Louis*
Nwakanma SIDNEY, *Washington University in St. Louis*
Soumendra LAHIRI, *Washington University in St. Louis*

In spatial regression, a challenging problem is selecting the correct set of non-zero covariates in many real-life studies. For example, in the election results, two closely located counties may have a similar set of relevant covariates, whereas for two distant counties, this similarity is very rare to observe. In the literature, researchers sought to address this problem using a grouped LASSO. Still, here the problem is that if a covariate is irrelevant in any county, that does not imply the covariate will be irrelevant in the entire spatial surface. In this context, we first formulate a local LASSO such that two closer locations will have a similar set of non-zero covariates. We capture the non-zero spatial signal in the coefficient function by the tensor product of the wavelet transformation in the prototype set of a spatial surface. As a result, if a spatial covariate surface is non-zero, then there will exist at least one non-zero signal in its prototype set in any resolution. We give the idea of how

to generalize local LASSO under the SCAD penalty. We theoretically validate the variable selection consistency in the increasing spatial domain under some regularity conditions. We have justified our results with two real-life election results.

234 . Gradient Equilibrium in Online Learning

[Plenary Lecture 3, (page 30)]

Ryan TIBSHIRANI, *University of California, Berkeley*
Anastasios ANGELOPOULOS, *University of California, Berkeley*
Michael JORDAN, *University of California, Berkeley*

We present a new perspective on online learning that we refer to as gradient equilibrium: a sequence of iterates achieves gradient equilibrium if the average of gradients of losses along the sequence converges to zero. In general, this condition is not implied by, nor implies, sublinear regret. It turns out that gradient equilibrium is achievable by standard online learning methods such as graadient descent and mirror descent with constant step sizes (rather than decaying step sizes, as is usually required for no regret). Further, as we show through examples, gradient equilibrium translates into an interpretable and meaningful property in online prediction problems spanning regression, classification, quantile estimation, and others. Notably, we show that the gradient equilibrium framework can be used to develop a debiasing scheme for black-box predictions under arbitrary distribution shift, based on simple post hoc online descent updates. We also show that post hoc gradient updates can be used to calibrate predicted quantiles under distribution shift, and that the framework leads to unbiased Elo scores for pairwise preference prediction.

235 . Nonparametric Empirical Bayes and Selective Inference

[01.E1.I15, (page 12)]

Surya TOKDAR, *Duke University*
Peter HOFF, *Duke University*

Consider a multi-population inference problem where it is of interest to estimate the mean of the population with the highest observed sample average. The usual confidence interval does not work in this case – offering increasingly lower coverage than the nominal value when the total number of populations gets larger. This phenomenon is often referred to as the Winner's Curse. Various modifications have been proposed to adjust for the selection step. We show that interval procedures that guarantee nominal coverage conditional on the selection event typically have infinite expected length. This result motivates us to consider empirical Bayesian solutions which offer coverage guarantees only on average over some parameter subspace. Nonparametric empirical Bayesian solutions are shown to generally offer good coverage with high precision but can perform poorly when one population is very different from all others – a clear violation of the underlying exchangeability assumption. We conclude with further mitigation strategies and discuss their frequentist and Bayesian interpretations.

236. 100 Years of Pies vs Bars

[03.A1.C2, (page 33)]

Maksuda Aktar TOMA, *University of Nebraska Lincoln*
Susan VANDERPLAS, *University of Nebraska Lincoln*

After William Playfair invented pie and bar charts in the early 1800s, textbooks and manuals expressed strong opinions regarding the effectiveness of other data displays around the turn of the century. Pie charts are "not a desirable form of presentation" (Brinton 1914) and "an insult to a man's intelligence" (Karsten 1923), according to Eells (1926). Although the "sector method" is prevalent, Brinton (1914) suggests using horizontal bar methods more often for speed and precision. These findings reflect today's situation: pie charts are still used despite many design guidelines and empirical studies identifying issues with circular data representations.

The Dawn of Empirical Graphics

The question of pies versus bars motivated some of the earliest empirical evaluations of charts, with studies asking participants to estimate quantities and measuring their accuracy and speed. Initial findings showed no clear preference between pie and bar charts; both were read with similar accuracy. However, the simplicity of these experiments drew criticism. Follow-up studies explored two-category charts using proportions like 0.25 and 0.4, with some researchers using A/B comparisons (where $A+B=1$) and others collecting estimates separately. This exchange marked the beginning of experimental statistical graphics as an area of interest for both researchers and practitioners.

Croxton and Stryker (1927) released more data on prediction accuracy using pie and bar charts, different categories, and section orientation/alignment. They found that pie charts outperformed bar charts in most cases, across scale alignments, and in charts with up to 5 categories (albeit differences were not statistically significant). Croxton and Stein (1932),

the final work in this early collection of comparisons, studied proportional data representations outside of statistical charts. The experimental stimuli, while applicable to charts, focus on general perception.

Modern Experimental Graphics

Much later, Cleveland and McGill (1984), a statistician and a psychologist by training, respectively, conducted experiments on simple graphical elements. These trials supported a graphical perceptual hierarchy that formed throughout the investigation. This rating suggests that bar charts are better viewed than pie charts because aligned length judgements are more accurate than area or angle judgements. Most of the hierarchy of feature comprehension is considered to be experimentally determined, whereas the Cleveland and McGill studies only inform a few aspects. As in the late 1920s, introspection-based reasoning and basic experiments were questioned. Psychologists Spence and Lewandowsky (1991) tested pie and bar charts for part-to-whole comparisons using forced-choice questions such "which is bigger, A or $B + C$".

Scope and Aims

In this study, we re-examine the historical literature surrounding the use of pie and bar charts, including heuristics and empirical studies. Where possible, we trace guidelines back to experimental studies and assess whether justifications based on those studies are directly supported by the evidence. We trace these forward in time, reviewing papers that cite them, summaries of the original studies, and other relevant information related to pie and bar charts. In addition, we specifically assess the different designs used in each experimental study, examining how the type of elicitation method impacts the results. Synthesizing these findings, we use the fundamental question of "Which is better, pies or bars?" to explore the broader history of experimental graphics and the evolution of graphical design guidelines.

237. Mixed Poisson families with real-valued mixing distributions

[02.A1.I31, (page 20)]

Will TOWNES, *Carnegie Mellon University*

Mixed Poisson families are widely used to model count data with overdispersion, zero inflation, or heavy tails in a variety of applications including finance, biology, and the physical sciences. The Poisson rate is typically assigned a nonnegative-valued mixing distribution. Surprisingly, it is also possible for the mixing distribution to have negative support. For example, the Hermite distribution is analogous to mixing a Poisson with a Gaussian and can be derived using generating functions so long as constraints on the natural parameter are satisfied. Here we provide general conditions on the mixing distribution that are necessary for a mixed Poisson to exist. A key tool is the use of subweibull bounds on the rates of tail decay and Lp norm growth. We illustrate the scope of mixed Poisson families with examples having different tail behaviors. Finally we comment on the mixed Poisson analogs of the skewed stable family.

238. Constructing confidence sequences from adaptive Robbins-Siegmund's lemma

[01.E1.I12, (page 11)]

Pham TUAN, *UT Austin*

TBA

239. TBD

[02.A1.I32, (page 20)]

Bingkai WANG, *Department of Biostatistics, University of Michigan*

TBA

240. The Role of Propensity Score in Leveraging External Data for Regulatory Decision-Making

[02.E1.I39, (page 23)]

CG WANG, *Regeneron*

Incorporating external data in regulatory decision-making requires much more than simply "mixing" external data with investigational clinical trial data. The external data must undergo appropriate analysis for deriving the right evidence for the investigational clinical trial. Moreover, such analysis must be integrated with the design and analysis of the investigational study for regulatory decision-making.

In this talk, we will review the different propensity score-integrated approaches for leveraging external data in clinical trial design and analysis, and clarify several aspects of these methods in the context of regulatory decision-making.

241. Improving the Efficiency of Clinical Trials by Making Efficacy Inferences Using Multivariate Endpoints Across Multiple Visits—Lesson Learned from DIAN-TU Platform Trial

[03.A1.I57, (page 31)]

Guoqiao WANG, *Washington University in St Louis*

Longitudinal data from clinical trials are com-

monly analyzed using mixed models for repeated measures (MMRM) when the time variable is categorical or linear mixed-effects models (i.e., random effects models) when the time variable is continuous. In these models, statistical inference is typically based on the absolute difference in the adjusted mean change (for categorical time) or the rate of change (for continuous time). Previously, we proposed a novel approach: modeling the percentage reduction in disease progression associated with the treatment relative to the placebo decline using proportional models. This concept of proportionality provides an innovative and flexible method for simultaneously modeling different cohorts, multivariate endpoints, and jointly modeling continuous and survival endpoints. A multivariate endpoint has been utilized as the primary endpoint for the DIAN-TU platform trial to improve its efficiency (i.e., reduce sample size). The objective of this presentation is to introduce the concept of this proportional model, justify its feasibility using real clinical trial examples, demonstrate its effectiveness relative to the test statistic of the difference between two means, and illustrate the increased power compared with traditional models.

242. Sequential Hypothesis Testing via No-Regret Learning

[03.A1.I59, (page 31)]

Jun-Kun WANG, *UCSD*
Can CHEN, *UCSD*

Online convex optimization (a.k.a. no-regret learning) concerns a scenario where an online learner commits to a point at each round before receiving the loss function. The learner's goal is to minimize the regret, defined as the gap between the cumulative losses and that of a clairvoyant who knows the sequence of the loss functions in advance. In this talk, I will first review a very neat known result in the literature that casts non-parametric sequential hypothesis testing as an online convex optimization problem, where an online learner tries to bet whether the null hypothesis is true or false, and a tighter regret bound suggests a faster stopping time to reject the null when the alternative is true. Then, I will show the relevant techniques can be used to design algorithms with strong statistical guarantees with applications such as online detecting LLM-generated texts, auditing fairness, and detecting distribution shifts. After that, I will introduce a new algorithm that overcomes the limitations of the existing methods and potentially leads to a faster rejection time under the alternative while controlling the false positive rate.

243. AI-Powered Surrogate Endpoint Validation for Oncology Trials: A Case Study in Multiple Myeloma

[01.A1.I5, (page 9)]

Lanjing WANG, *University of Washington*
Zixuan ZHAO, *George Washington University*
Zexin REN, *George Washington University*
Will MA, *Hope AI*

Survival endpoints in oncology and hematology clinical trials often delay market access due to extended follow-up requirements. Surrogate endpoints can accelerate approval, but their validation traditionally requires time-consuming literature reviews and multi-institutional data sharing. We developed an AI-powered workflow to rapidly synthesize clinical evidence from relevant studies and generate synthetic individual patient data (IPD) through digitization of Kaplan-Meier plots, preserving survival time and covariate information. Application to multiple myeloma validates association between minimal residual disease negativity (MRD-) and median progression-free survival (mPFS) at both trial and patient levels. Our expert-supervised AI approach reduced analysis time from the original 5 years to under two weeks, maintaining comprehensive and robust statistical rigor.

244. High-dimensional Change-point Detection Using Generalized Homogeneity Metrics

[03.M2.I53, (page 29)]

Runmin WANG, *Texas A&M University*
Shubhadeep CHAKRABORTY, *Bristol Myers Squibb Company*
Xianyang ZHANG, *Texas A&M University*

In this talk, we study the problem of detecting abrupt changes in the data-generating distributions of a sequence of high-dimensional observations beyond the first two moments. This problem has remained substantially less explored in the existing literature, especially in the high-dimensional context, compared to detecting changes in the mean or the covariance structure. We develop a nonparametric methodology to (i) test the existence of a change-point, and (ii) identify the change-point locations in an independent sequence of high-dimensional observations. Our approach rests upon recent nonparametric tests for the homogeneity of two high-dimensional distributions. We construct a single change-point test statistic based on a cumulative sum process in an embedded Hilbert space. We shall derive its limiting null distribution and present the asymptotic consistency under the high dimension

medium sample size framework. We also combine our statistics with wild binary segmentation to recursively estimate and test for multiple change-point locations. The superior performance of our methodology compared to other existing procedures will be illustrated via extensive simulation studies and the application to the stock return data observed during the period of the global financial crisis in the United States.

245. Microbiome Data Integration via Shared Dictionary Learning

[04.M1.I71, (page 36)]

Shulei WANG, *University of Illinois Urbana-Champaign*
Bo YUAN, *University of Illinois Urbana-Champaign*

Data integration is a powerful tool for facilitating a comprehensive and generalizable understanding of microbial communities and their association with outcomes of interest. However, integrating data sets from different studies remains a challenging problem because of severe batch effects, unobserved confounding variables, and high heterogeneity across data sets. We propose a new data integration method called MetaDICT, which initially estimates the batch effects by weighting methods in causal inference literature and then refines the estimation via novel shared dictionary learning. Compared with existing methods, MetaDICT can better avoid the overcorrection of batch effects and preserve biological variation when there exist unobserved confounding variables, data sets are highly heterogeneous across studies, or the batch is completely confounded with some covariates. Furthermore, MetaDICT can generate comparable embedding at both taxa and sample levels that can be used to unravel the hidden structure of the integrated data and improve the integrative analysis. Applications to synthetic and real microbiome data sets demonstrate the robustness and effectiveness of MetaDICT in integrative analysis. Using MetaDICT, we characterize microbial interaction, identify generalizable microbial signatures, and enhance the accuracy of outcome prediction in two real integrative studies, including an integrative analysis of colorectal cancer metagenomics studies and a meta-analysis of immunotherapy microbiome studies.

246. Sparse Autoencoders Demystified: Provable Feature Learning via Adaptive Bias Scheduling

[03.M2.I55, (page 29)]

Tianhao WANG, *Toyota Technological Institute at Chicago*

Sparse Autoencoders (SAEs) are a popular tool for interpreting the learned features in trained deep neural networks. However, SAEs are notoriously difficult to train due to the requirement of sparse activation patterns, and it still lacks a principled understanding of how SAEs can learn the true features when they are mixed in the data. In this work, we demystify the training of SAEs by providing a theoretical characterization of the learning process, which further inspires a novel method for training SAEs based on adaptive bias scheduling. We illustrate that our method outperforms the L1 regularization-based training for SAEs on synthetic tasks as well as on transformers trained on modular arithmetic tasks. Moreover, we propose a knowledge-stitching technique for the transformer model, which integrates the attention layer with the learned features from the SAE. We empirically show that the knowledge-stitching technique not only enhances the interpretability of the learned features but also improves the quality of the learned features.

247. Long-term causal inference under persistent confounding via data combination

[01.A1.I6, (page 9)]

Yuhao WANG, *Tsinghua University*
Guido IMBENS, *Stanford University*
Nathan KALLUS, *Cornell University*
Xiaojie MAO, *Tsinghua University*

We study the identification and estimation of long-term treatment effects by combining short-term experimental data and long-term observational data subject to unobserved confounding. This problem arises often when concerned with long-term treatment effects since experiments are often short-term due to operational necessity while observational data can be more easily collected over longer time frames but may be subject to confounding. In this paper, we tackle the challenge of persistent confounding: unobserved confounders that can simultaneously affect the treatment, short-term outcomes, and long-term outcome. In particular, persistent confounding invalidates identification strategies in previous approaches to this problem. To address this challenge, we exploit the sequential structure of multiple short-term outcomes and develop several novel identification strategies for the average long-term treatment effect. Based on these, we develop estimation and inference methods with asymptotic guarantees. To demonstrate the importance of handling persistent confounders, we apply our methods to estimate

the effect of a job training program on long-term employment using semi-synthetic data. This paper has been accepted for publication at JRSSB, DOI: 10.1093/jrsssb/qkae095.

248. How environmental clustering revealed confusion in the statistical literature and how we fixed it

[01.A1.I9, (page 10)]

Kevin WRIGHT, *Corteva Agriscience*

Corteva tests new crop varieties in field trials at thousands of locations around the world. Grouping the location environments into clusters can help us understand where the new varieties perform best. During our research on the use of Hopkins Statistic for clustering environmental covariates, we found different definitions for the statistic. We identified the correct definition of the statistic and created a new R package to correctly and efficiently calculate Hopkins Statistic.

249. Recent theoretical advances in diffusion models

[02.M2.I29, (page 19)]

Yuchen WU, *University of Pennsylvania*
Andrea MONTANARI, *Stanford University*
Yuxin CHEN, *University of Pennsylvania*
Yuting WEI, *University of Pennsylvania*

In this talk, I will present recent theoretical advances in diffusion models, a class of deep generative models driving many cutting-edge applications. In the first part, I will introduce a training-free acceleration method for diffusion models. Our approach is simple to implement, compatible with any pre-trained diffusion model, and comes with a convergence rate that strengthens prior theoretical results. We demonstrate the effectiveness of our algorithm across multiple real-world image generation tasks. In the second part, I will discuss a new class of sampling algorithms designed based on the structure of diffusion models. Our approach replaces score networks in the diffusion model architecture with more efficient denoising algorithms that encode information about the target distribution. As applications, we use our method for posterior sampling in two high-dimensional statistical problems: sparse regression and low-rank matrix estimation within the spiked model. In both cases, we develop algorithms with accuracy guarantees in the regime of constant signal-to-noise ratios.

250. Tree-Regularized Bayesian Latent Class Analysis for Improving Weakly Separated Dietary Pattern Subtyping in Small-Sized Subpopulations

[02.M1.I25, (page 17)]

Zhenke WU, *Department of Biostatistics, University of Michigan*
Mengbing LI, *University of Michigan*
Briana STEPHENSON, *Harvard University*

Dietary patterns synthesize multiple related diet components, which can be used by nutrition researchers to examine diet-disease relationships. Latent class models (LCMs) have been used to derive dietary patterns from dietary intake assessment, where each class profile represents the probabilities of exposure to a set of diet components. However, LCM-derived dietary patterns can exhibit strong similarities, or weak separation, resulting in numerical and inferential instabilities that challenge scientific interpretation. This issue is exacerbated in small-sized subpopulations. To address these issues, we provide a simple solution that empowers LCMs to improve dietary pattern estimation. We develop a tree-regularized Bayesian LCM that shares statistical strength between dietary patterns to make better estimates using limited data. This is achieved via a Dirichlet diffusion tree process that specifies a prior distribution for the unknown tree over classes. Dietary patterns that share proximity to one another in the tree are shrunk towards ancestral dietary patterns a priori, with the degree of shrinkage varying across pre-specified food groups. Using dietary intake data from the Hispanic Community Health Study/Study of Latinos, we apply the proposed approach to a sample of 496 US adults of South American ethnic background to identify and compare dietary patterns.

251. Bayesian Nonparametric Methods for Oncology Studies: Riten Mitra's Impact in Cancer Research

[02.M1.I20, (page 15)]

Yanxun XU, *Associate Professor, Department of applied mathematics and statistics, Johns Hopkins University*

TBA

252. Elastic Net-Based Variable Selection for Fréchet Regression in RKHS

[03.E1.I68, (page 35)]

Haoyi YANG, *Department of Statistics, The Pennsylvania State University, USA*
Bing LI,
Lingzhou XUE,
Satarupa BHATTACHARJEE,

We introduce a new framework for sparse Fréchet regression where both the response and covariates lie in general metric spaces. Unlike previous work, we focus on $d^2(Y, y\_1) - d^2(Y, y\_2)$ for fixed (y_1, y_2), treating it as the target of regression. This formulation preserves an additive structure in Hilbert spaces and naturally extends to general metric spaces.

Building upon this structure, we propose an additive model with elastic net regularization in reproducing kernel Hilbert spaces (RKHS) to enable variable selection. A key feature of our method is the sparsity-invariant property: the active set remains stable across a wide range of choices for (y_1, y_2), providing a robust and tuning-insensitive approach to variable selection.

We establish theoretical guarantees for our estimator, including the KKT conditions and variable selection consistency. Simulation studies based on data splitting confirm the stability of the selection procedure and provide strong empirical evidence supporting the sparsity invariance property in both distributional and SPD matrix regression settings.

253 . Improved automated cryo-EM structure determination via diffusion model

[04.M1.I71, (page 37)]

Yisha YAO, *Columbia University*
Xiaoyu FANG, *Columbia University*
Sheng CHEN, *Sun Yat-sen University*

Cryo-electron microscopy (cryo-EM) is a revolutionary technique to visualize biological molecules and determine their three-dimensional structures. It allows studying large, flexible, and heterogeneous macromolecular complexes in their native status, and captures molecular dynamics in one single experiment. To push forward these frontiers, the key lies in building accurate atomic 3D models on the experimental EM density maps. Typically, the peptide chain(s) is known, and the task boils down to identify the position of all the atoms for the peptide chain(s) in the cryo-EM density map. The traditional solution for this task involves manual searching for templates, visual inspection of the 3D structure, and model refinement. Such procedure depends on human expertise and is time-consuming. To improve throughput, reproducibility, and accessibility as well as unlock its full potential, it is imperative to develop automated software for cryo-EM structure determination. In this work, we integrate diffusion-model-based denoising techniques to predict amino acid types at each 3D position on the EM density map. Combined with C atoms prediction and backbone atoms prediction through multi-task deep learning, we are able to build accurate 3D atomic structure models. The results show some improvement compared to existing methods.

254. The Evolution of Crops County Estimates: Past, Present and Future

[Plenary Lecture 2, (page 20)]

Linda YOUNG,

255. Expectation Maximization Estimation for Hawkes Process with Missingness

[02.M1.C1, (page 16)]

Jingtian YU, *Oregon State University*
Jingtian YU, *Oregon State university*
Sharmodeep BHATTACHARYYA, *Oregon State university*
Sarah EMERSON, *Oregon State university*

The Hawkes process is a self-exciting point process for modeling point process data in a diverse range of fields, including finance, neuroscience, and social networks. The occurrence of missing data is quite common in point process data. The event intensity of the Hawkes process depends on past occurrences, making estimation with incomplete observations a challenge as unobserved events affect the intensity function. Much past work on estimation of Hawkes processes with missing data focuses on scenarios in which the missing events can be confined to a known interval, typically non-overlapping with the observation window or a subset of the observation window. The more general scenario in which missing events can occur at any point within the observation window remains unexplored. In this work, we describe a general mechanism of missingness where the probability of an event being missing does not depend on the event's position in the time sequence. We develop a likelihood-based estimation approach that incorporates imputation steps tailored to accommodate the missing mechanism, ensuring better handling of estimation bias in the incomplete data scenarios. We provide an extensive simulation study demonstrating the superior performance of our proposed estimation method. We also provide examples in real-life event time datasets.

256. Test-negative designs with various reasons for testing: statistical bias and solution

[01.E1.I14, (page 12)]

Mengxin YU, *University of Pennsylvania*

Test-negative designs are widely used for post-market evaluation of vaccine effectiveness, particularly in cases when randomized trials are not feasible. Differing from classical test-negative designs where only healthcare-seekers with symptoms are included, recent test-negative designs have involved individuals with various reasons for testing, especially in an outbreak setting. While including these data can increase sample size and hence improve precision, concerns have been raised about whether they introduce bias into the current framework of test-negative designs, thereby demanding a formal statistical examination of this modified design. In this article, using statistical derivations, causal graphs, and numerical demonstrations, we show that the standard odds ratio estimator may be biased if various reasons for testing are not accounted for. To eliminate this bias, we identify three categories of reasons for testing, including symptoms, mandatory screening, and case contact tracing, and characterize associated statistical properties and estimands. Based on our characterization, we show how to consistently estimate each estimand via stratification. Furthermore, we describe when these estimands correspond to the same vaccine effectiveness parameter, and, when appropriate, propose a stratified estimator that can incorporate multiple reasons for testing and improve precision. The performance of our proposed method is demonstrated through simulation studies.

257 . Bayesian Federated Cause-of-Death Quantification Under Distribution Shift

[02.M1.I25, (page 17)]

Li ZEHANG, *University of California, Santa Cruz*
Yu ZHU, *University of California, Santa Cruz*

Cause-of-death data is fundamental for understanding population health trends and inequalities as well as designing and evaluating public health interventions. A significant proportion of global deaths, particularly in low- and middle-income countries (LMICs), do not have medically certified causes assigned. In such settings, verbal autopsy (VA) is a widely adopted approach to estimate disease burdens by interviewing caregivers of the deceased. In this talk, we propose a flexible Bayesian federated learning approach that enables cause-of-death assignment and quantification of cause-of-death distribution in a new population without data sharing from multiple training datasets. The key to our approach is a latent class model framework that allows flexible characterization of the joint distribution of symptoms and causes across heterogeneous populations with distribution shift. We show that the proposed method significantly outperforms models based on a single training dataset and achieves comparable performance compared to joint modeling approach that pools all available data. We will also discuss practical implications of such federated learning models in VA analysis pipeline.

258. Joint analysis for multivariate longitudinal and interval-censored event time data: Application in Huntington's disease

[01.M2.I4, (page 4)]

Yue ZHAN, *University of Nebraska Medical Center*
Cheng ZHENG, *University of Nebraska Medical Center*
Ying ZHANG, *University of Nebraska Medical Center*

We develop a joint model of multivariate longitudinal biomarkers with a change point at an interval-censored event time. Our model allows us to simultaneously understand the causal effect of longitudinal biomarkers on the event time and the causal effect of event time on the changes of longitudinal biomarkers post the event. A simulation study is carried out to demonstrate the satisfactory finite-sample performance of the proposed method for making inferences. Finally, the method is applied to PREDICT-HD data from a multisite observational cohort study of prodromal Huntington's disease individuals to ascertain the effects of cognitive impairments on the onset of Huntington's disease that subjects to interval censoring and how the disease onset accelerates the cognitive impairments.

259. GAE-BEG Model: A novel GNN Framework integrating neuroimaging and behavioral information to understand Adolescent Psychiatric Disorders

[01.E1.I13, (page 12)]

Aiying ZHANG, *University of Virginia*
Gang QU, *Tulane University*

Functional connectivity (FC) provides insights into multiple psychiatric disorders, yet the substantial inter-subject variability of FC hinders its effectiveness in distinguishing between various psychiatric disorders. Therefore, we propose a novel graph learning framework that integrates FC with behavioral characteristics to better differentiate between psychiatric disorders. Additionally, applying Grad-CAM enhances model interpretability by identifying key regions of interest (ROIs) involved in distinguishing individuals with psychiatric disorders from those without. Preliminary experiments using

the ABCD dataset revealed two key findings: first, critical regions including the thalamus, putamen, and pallidum, along with nodes from the somatomotor and cingulo-opercular networks, are essential for distinguishing psychiatric disorders. Additionally, visualization of latent representations indicated that individuals with externalizing disorders, specifically Oppositional Defiant Disorder (ODD), are notably distinguishable from healthy controls. These results highlight the potential of our graph learning framework in identifying psychiatric disorders.

260. High dimensional mediation analysis with applications in genetics

[03.M1.I52, (page 28)]

Qi ZHANG, *University of New Hampshire*
Qi ZHANG, *University of New Hampshire*

To leverage the advancements in GWAS and QTL mapping for traits and molecular phenotypes to gain mechanistic understanding of the genetic regulation, biological researchers often investigate the eQTLs that colocalize with QTL or GWAS peaks. Our research is inspired by two such studies. One aims to identify the causal SNPs that are responsible for the phenotypic variation and whose effects can be explained by their impacts at the transcriptomic level in maize. The other study in mouse focuses on uncovering the cis-driver genes that induce phenotypic changes by regulating trans-regulated genes. Both studies can be formulated as mediation problems with potentially high- dimensional exposures, confounders and mediators that seek to estimate the overall indirect effect for each exposure. In this paper, we propose MedDiC, a novel procedure to estimate the overall indirect effect based on difference- in-coeffcients approach. Our simulation studies nd that MedDiC others valid inference for the indirect effect with higher power, shorter condence intervals and faster computing time than competing methods. We apply MedDiC to the two aforementioned motivating datasets, and nd that MedDiC yields reproducible outputs across the analysis of closely related traits, with results supported by external biological evidence. The code and additional information are available on our Github page (https://github.com/QiZhangStat/MedDiC).

261. Distance and Kernel-Based Measures for Global and Local Two-Sample Conditional Distribution Testing

[02.E1.I41, (page 23)]

Xianyang ZHANG, *Texas A&M University*
Jian YAN, *Cornell University*
Zhuoxi LI, *Xiamen University*

Testing the equality of two conditional distributions is crucial in various modern applications, including transfer learning and causal inference. Despite its importance, this fundamental problem has received surprisingly little attention in the literature. This work aims to present a unified framework based on distance and kernel methods for both global and local two-sample conditional distribution testing. To this end, we introduce distance and kernel-based measures that characterize the homogeneity of two conditional distributions. Drawing from the concept of conditional U-statistics, we propose consistent estimators for these measures. Theoretically, we derive the convergence rates and the asymptotic distributions of the estimators under both the null and alternative hypotheses. Utilizing these measures, along with a local bootstrap approach, we develop global and local tests that can detect discrepancies between two conditional distributions at global and local levels, respectively. Our tests demonstrate reliable performance through simulations and real data analyses.

262. A model-agnostic ensemble framework with built-in LOCO feature importance inference

[01.M2.I1, (page 3)]

Lili ZHENG, *University of Illinois Urbana - Champaign*
Luqin GAN,
Genevera ALLEN, *Columbia University*

Interpretability and reliability are crucial desiderata when machine learning is applied in critical applications. However, generating interpretations and uncertainty quantifications for black-box ML models often costs significant extra computation and held-out data. In this talk, I will introduce a novel ensemble framework where one can simultaneously train a predictive model and gives uncertainty quantification for its interpretation, in the form of leave-one-covariate-out (LOCO) feature importance. This framework is almost model-agnostic, can be applied with any base model, for regression or classification tasks. Most notably, it avoids model-refitting and data-splitting, and hence there is no extra cost, computationally and statistically, for uncertainty quantification. To ensure the inference validity without data splitting, we address a number of challenges by leveraging the stability of the ensemble training process. I will discuss some broad connection of this work to selective inference, and other model-agnostic

feature importance inference methods. I will also demonstrate the framework via some real benchmark datasets.

263. A statistical theory of overfitting for imbalanced classification

[03.A1.I61, (page 32)]

Kangjie ZHOU, *Columbia University*
Jingyang LYU, *University of Wisconsin, Madison*
Yiqiao ZHONG, *University of Wisconsin, Madison*

Classification with imbalanced data is a common challenge in data analysis, where certain classes (minority classes) account for a small fraction of the training data compared with other classes (majority classes). Classical statistical theory based on large-sample asymptotics and finite-sample corrections is often ineffective for high-dimensional data, leaving many overfitting phenomena in empirical machine learning unexplained.

In this talk, we develop a statistical theory for high-dimensional imbalanced classification by investigating support vector machines and logistic regression. We find that dimensionality induces truncation or skewing effects on the logit distribution, which we characterize via a variational problem under high-dimensional asymptotics. In particular, for linearly separable data generated from a two-component Gaussian mixture model, the logits from each class follow a normal distribution (0,1) on the testing set, but asymptotically follow a rectified normal distribution max , (0,1) on the training set – which is a pervasive phenomenon we verified on tabular data, image data, and text data. This phenomenon explains why the minority class is more severely affected by overfitting. Further, we show that margin rebalancing, which incorporates class sizes into the loss function, is crucial for mitigating the accuracy drop for the minority class. Our theory also provides insights into the effects of overfitting on calibration and other uncertain quantification measures.

264 . Addressing Antidiscrimination with Variational Inference

[03.E1.I65, (page 34)]

Shuang ZHOU, *Arizona State University*
Lydia GABRIC, *Arizona State University*
Kenneth ZHOU, *University of Waterloo*

Within the insurance industry, evolving antidiscrimination regulations have led insurers to exclude protected information in pricing calculations. With the rise of complex algorithmic methods and big data, insurers face the dual challenge in complying with regulatory standards while maintaining reasonable estimates. Many researchers have proposed methods to achieve discrimination-free pricing and varying fairness measures, but these techniques often require protected information at the individual-level which is often inaccessible. In this work, we propose to study the indirect discrimination via a hierarchical finite mixture model with latent protected variables. The hierarchical structure of the model will be represented through the prior specification, and the posterior distribution is used to impute the unobserved sensitive variable through the indirect discriminatory dependence structure. To tackle the indirect discrimination existing in the true posterior, we propose the use of variational inference with a mean-field variational family that enforces independence between unknown variables, eliminating the indirect discrimination by definition. A further step of importance sampling is incorporated to achieve insurance unbiasedness with the optimized discrimination-free distribution. Our method is supported with a simulation study inspired by real insurance data.

265. High-dimensional moderated mediation analysis with heredity

[04.M1.I70, (page 36)]

Wen ZHOU, *New York University*
Zhang ZIFENG, *Colorado State University*
Fan YANG, *Tsinghua University*
Peng DING, *UC Berkeley*

In recent years, there has been a significant increase in attention to high-dimensional mediation inference in various fields, including economics, finance, and genomic and genetic research. A key challenge in this domain is the inference of natural direct and indirect effects in the presence of potential interactions between treatment and high-dimensional mediators. These interactions often give rise to moderator effects, which are further complicated by the intricate dependencies among the mediators. In this paper, we introduce a new inference procedure that addresses this challenge. By incorporating a nonconvex penalty into the outcome model, our method effectively identifies important mediators while accounting for their interactions with the treatments, which admits the guaranteed oracle property. Leveraging the oracle property, we can exploit a projection onto the mediator model, guided by the estimated important direction in the mediator space. We establish the asymptotic normality of both natural indirect and direct effects for inference. Additionally, we develop an algorithm that utilizes the overlapped group smoothly clipped absolute deviation penalty to promote heredity structure among the main effects

and interactions, which comes with provable guarantees. Our extensive numerical studies, comparing our method with other existing approaches across various scenarios, demonstrate its effectiveness. To illustrate the practical application of our methods, we conduct a study investigating the impact of childhood trauma on cortisol stress reactivity. Using DNA methylation loci as mediators, we uncover several new loci that remain undetected when interactions are ignored.

266 . Testing independence for sparse longitudinal data

[02.E1.I38, (page 23)]

Changbo ZHU, *University of Notre Dame*
Wang JANE-LING,

With the advance of science and technology, more and more data are collected in the form of functions. A fundamental question for a pair of random functions is to test whether they are independent. This problem becomes quite challenging when the random trajectories are sampled irregularly and sparsely for each subject. In other words, each random function is only sampled at a few time-points, and these time-points vary with subjects. Furthermore, the observed data may contain noise. To the best of our knowledge, there exists no consistent test in the literature to test the independence of sparsely observed functional data. We show in this work that testing pointwise independence simultaneously is feasible. The test statistics are constructed by integrating pointwise distance covariances (Székely et al., 2007) and are shown to converge, at a certain rate, to their corresponding population counterparts, which characterize the simultaneous pointwise independence of two random functions. The performance of the proposed methods is further verified by Monte Carlo simulations and analysis of real data.

Directory

Local Map

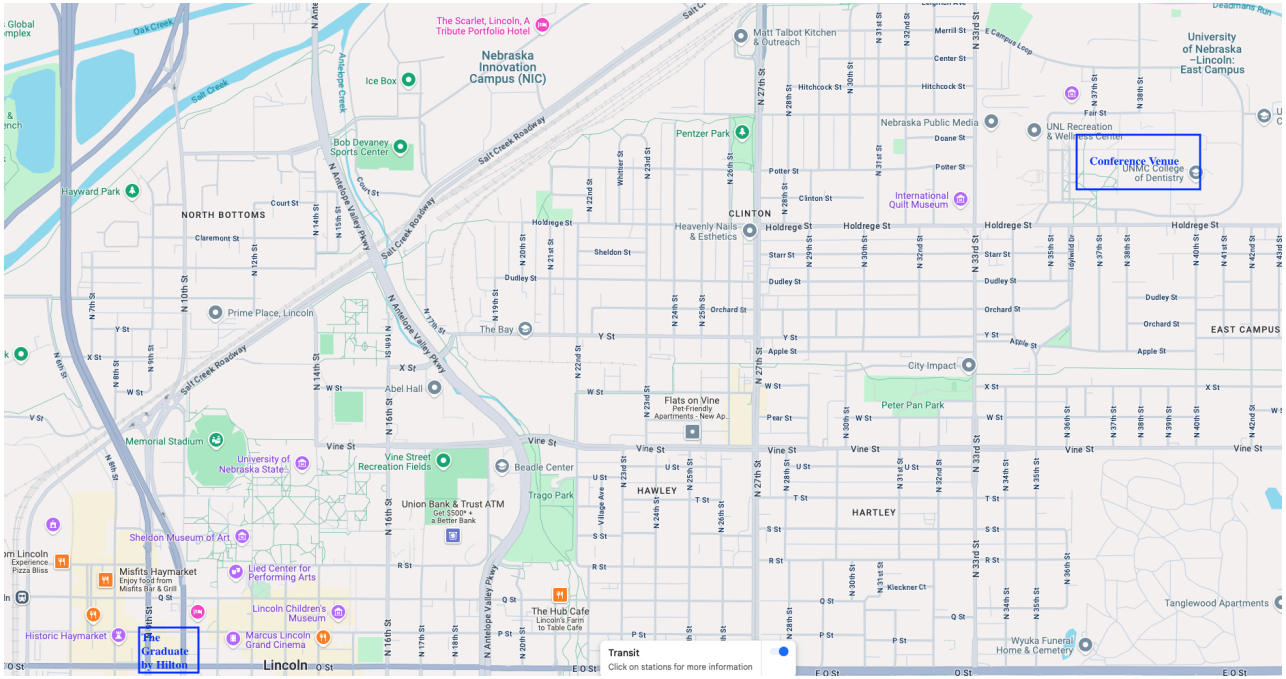


Figure 1: The map of the Downtown Lincoln and the East Campus

- All sessions will be held at the Nebraska East Union (NEU) at the East campus of the University of Nebraska-Lincoln (UNL).
- The venue is about 3.9 miles away from the conference hotel, i.e., The Graduate by Hilton, which is located in the Haymarket area of Downtown Lincoln..
- The hotel is next to the city campus of UNL, a different campus from the East campus where the conference is being held.
- From the hotel, NEU is about 9 minutes by car—around 20 minutes by public bus.
- The public buses 24, 25 (weekdays only), 42, and 49 (weekdays and Saturday) run from the downtown area/city campus to the east campus.
- Ride-hailing services like Uber are usually readily available.

East Campus

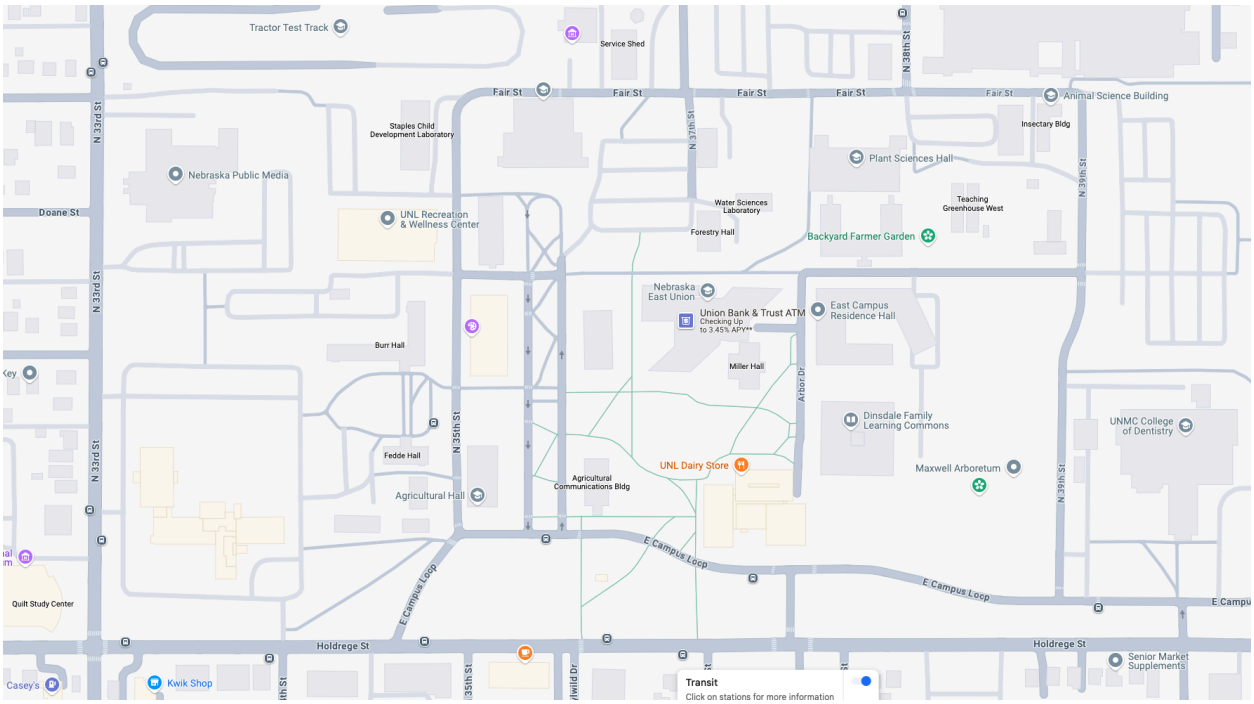


Figure 2: Map of the UNL East Campus, NEU ie. the conference venue and the Messengale residential Complex is located.

- The Miller Hall does not exist.
- The buses 24 and 25 run through the East Campus Loop. The nearest stop is in front of the UNL dairy store.
- These buses circle between the city and the east campus.
- The bus 42 travels on the Holdredge Street. The nearest stop is Holdredge and $37th$.
- The bus 49 travels through the $33rd$ street. The nearest stop is on the $33rd$ street opposite the Quilt Museum.
- The closest bus stands to The Graduate would be on P Street, either behind the Manse building or opposite to the Embassy Suites hotel.
- The buses accept fare ($1.25) on board (no change), or you can buy a bus pass on the StarTran website.

The Graduate, Haymarket, and City campus