



2022 ICSA
APPLIED STATISTICS
SYMPOSIUM

Gainesville, Florida
June 19-22, 2022



International Chinese Statistical Association

泛華統計協會

2022 ICSA APPLIED STATISTICS SYMPOSIUM
CONFERENCE INFORMATION, PROGRAM AND ABSTRACTS

June 19 - 22, 2022
University of Florida
Gainesville, Florida, USA

Organized by
International Chinese Statistical Association

Hosted by
Department of Biostatistics, University of Florida

Contents

Welcome	1
Conference Information	2
ICSA Officers and Committees	2
Conference Committees	9
Sponsor Information	13
Transportation and Parking	27
Map	28
Conference Venue Information	29
Dining and WiFi Information	31
Program Overview	34
Keynote Lectures	35
Student Paper Awards	41
Short Courses	42
Scientific Program	47
Welcome and Opening Remarks: Mon, June 20 8:00-8:30 (EDT)	47
Plenary Keynote Talk 1: Mon, June 20 8:30-9:30 (EDT)	47
Sessions 1A-1H: Mon, June 20 10:00-11:40 (EDT)	47
Sessions 2B-2H: Mon, June 20 13:00-14:40 (EDT)	49
Sessions 3A-3H: Mon, June 20, 15:00-16:40 (EDT)	50
Sessions 4A-4H: Mon, June 20, 17:00-18:40 (EDT)	52
Plenary Keynote Talk 2: Tue, June 21, 8:30-9:30 (EDT)	53
Sessions 5A-5H: Tue, June 21, 10:00-11:40 (EDT)	54
Special Invited Talks: Tue, June 21, 13:00-14:30 (EDT)	55
Sessions 6A-6H: Tue, June 21, 15:00-16:40 (EDT)	55
Sessions 7A-7H: Tue, June 21, 17:00-18:40 (EDT)	57
Banquet Talk: Tue, June 21, 20:00-20:45 (EDT)	59
Plenary Keynote Talk 3: Wed, June 22, 8:30-9:30 (EDT)	59
Sessions 8A-8H: Wed, June 22, 10:00-11:40 (EDT)	59
Sessions 9A-9H: Wed, June 22, 13:00-14:40 (EDT)	60
Posters	63
Abstracts	65
Session 1A : Causal Inference And Its Applications	65
Session 1B : Latent Variable Models In The Data Science Era	65
Session 1C : Some Advances In Statistical Machine Learning	66
Session 1D : Machine Learning/Artificial Intelligence In Biomedical Research With 'big' Data	67
Session 1E : Statistical Challenges And Advances In Complex Data Analysis	68
Session 1F : Statistical Methods And Applications For Analyzing Real-World Data	68
Session 1G : Recent Advances In Survival And Recurrent Events Analysis For Complex Data Structures	69
Session 1H : Statistical Inference For Two-Phase Studies With Outcome-Dependent Sampling	70
Session 2B : Advanced Research In Bio-Molecular And Imaging Data By Our Young Researchers	71
Session 2C : Emerging Topics In Statistical Learning For Biomedical Data	72
Session 2D : Statistics In Biosciences (Sibs): Real World Challenges And Recent Methodological Developments	72
Session 2E : Some Recent Methods For Sequential Monitoring Of Complex Data	73
Session 2F : Big Data, Machine Learning And Graphical Methods	74
Session 2G : Recent Development In Survival Analysis In Clinical Trials	75

Session 2H : Challenges And Recent Developments In Multi-Outcome Analysis	76
Session 3A : Recent Advances In Statistical Methods For Causal Inference And Personalized Medicine	77
Session 3B : New Advances In High-Dimensional Data Analysis	78
Session 3C : Machine Learning And Deep Learning Methods For Complex And Big Data	79
Session 3D : Advance In Statistical Methods For Complex Data	80
Session 3E : Recent Advancement In Statistical Learning Methods For High-Dimensional Biomedical Data	80
Session 3F : Advanced Statistical Learning Methods For Dynamic Systems	81
Session 3G : Geometric Statistics In Medical Image Computing	82
Session 3H : The Jiann-Ping Hsu Invited Session on Biostatistical and Regulatory Sciences	83
Session 4A : Recent Developments For Causal Inference: Theory, Method, And Application	84
Session 4B : High-Dimensional Statistical Inference For Big Complicated Data	84
Session 4C : New Fronts In Joint Modeling And Machine Learning	85
Session 4D : Knowledge-Guided Machine Learning And Statistical Modeling In Longitudinal Studies With Survival Endpoints	86
Session 4E : Robust Information Integration From Multiple Studies In Clinical And Biomedical Research	87
Session 4F : Statistical Innovation In Complex And High Dimensional Data	88
Session 4G : Enhance Decision Making In Early Oncology Studies To Expedite Drug Development	88
Session 4H : Design And Analysis Of Computer Experiments	89
Session 5A : Statistical Methodologies In Causal Inference With Application In Drug Development	89
Session 5B : Recent Developments Of Dimension Reduction In Integrating Big And Complex Data	90
Session 5C : Precision Digital Health Care Via Machine Learning	91
Session 5D : Statistical Methods For Complex And High Dimensional Data	92
Session 5E : Modern Streaming Data Analysis: Change-Point Problems And Applications	93
Session 5F : Emerging Development In The Analysis Of Data With Complex Features	93
Session 5G : Statistical Leadership In Drug Development In The New Era Of Data Science	94
Session 5H : Student Paper Competition Winners	95
Session 6A : Recent Advances In Mendelian Randomization	96
Session 6B : Recent Advances In Dimension Reduction Techniques	97
Session 6C : Statistical Methods For Assessing Genomic Heterogeneity	97
Session 6D : Novel Statistical Modeling And Computing Methods For Complex Data	98
Session 6E : Modern Streaming Data Analysis: Detection And Identification	99
Session 6F : Deep Learning With Application And Uncertainty Quantification	99
Session 6G : Recent Advances In Clinical Trial Design And Practice	100
Session 6H : New Developments In Modern Nonparametric Statistics And The Applications	101
Session 7A : Novel Statistical Methods For -Omic Data Analysis	102
Session 7B : Modern Time Series And Network Methods In Data Science.	103
Session 7C : Innovative Approach Of Hidden Markov Model	103
Session 7D : Statistical Advances And Applications In Analyzing Large Scale & Multi-Omic Single-Cell Data	104
Session 7E : Modern Streaming Data Analysis: Process Monitoring	105
Session 7F : Discriminant And Cluster Analysis For Complex Data	106
Session 7G : Design And Analysis In Vaccine Development And Its Challenges	106
Session 7H : Methods For Inference On Variable Importance Using Machine Learning	107
Session 8A : Ultra-High Dimensional Variable Selection And Zero-Inflated Negative Binomial Spatial And Temporal Regression	108
Session 8B : Recent Developments In Functional Data Analysis	109
Session 8C : Recent Advances In Robust Statistical Models For Censored And Missing Data	110
Session 8D : Recent Advances In Latent Variable Analysis	110
Session 8E : New Advances In Microbiome Related Data Analysis	111
Session 8F : Statistical Computation Of Big Data With Biomedical Applications	112
Session 8G : Recent Development In Innovative Clinical Trial Designs	113
Session 8H : Some Popular Applications In Data Integration	113
Session 9A : Bayesian Calibration Of Computer Models	114
Session 9B : Novel Developments For Functional Data Analysis	115
Session 9C : Statistical Methods For High Dimensional Microbiome Data	115
Session 9D : Recent Advancements In Statistical Data Integration	116
Session 9E : Modern Business Statistical Analysis	117

Session 9F : Application And Theory Of Statistical Test And Evaluation	118
Session 9G : Statistical Challenges In Clinical Trials For Alzheimer Disease	119
Session 9H : Statistics Education In The Era Of Ai And Data Science	119
Index of Authors	120

2022 ICSA Applied Statistics Symposium

June 19-22, 2022

Gainesville, Florida, USA

On behalf of the organizing committee, we welcome you to the campus of the University of Florida. We are thankful to the ICSA executive committee for selecting Gainesville, FL, to be the venue for the 2022 Applied Statistics Symposium. We are super excited to have the unique distinction of hosting this conference face to face since the beginning of the pandemic. We plan to follow all safety protocols to create a safe environment for our participants.

The program committee has worked diligently to bring you an exciting program which consists of 3 plenary and 2 special invited lectures, plus 70 invited sessions on the theme of "Statistical Innovations in the Era of Artificial Intelligence and Data Science". In addition, there are 36 posters, 6 short courses and 5 oral presentations by student paper award recipients, and an after-dinner talk. In addition to the academic components of the conference, the local organizing committee has organized several social events including the opening mixture, entertainment programs, local outings, and a fabulous banquette.

We hope you will enjoy the next three and half days on campus. Besides attending numerous technical sessions, do take advantage of various local attractions Gainesville has to offer including a historic downtown, and various state and city parks with extensive nature trails. UF displays a very impressive campus listed on the National Register of Historic Places. You will be able to view beautiful brick construction buildings that showcase traditional Gothic architecture. A number of natural springs are within driving distance and so are three major metropolis and numerous beaches on both sides of the state. Feel free to contact the local organizers or the student volunteers if you need any assistance.

We thank all individuals and entities who contribute to the success of this event, notably, the deans of our two colleges, chair of department of biostatistics at UF, the staff members, the student volunteers, all committee members, our sponsors, notably the National Science Foundation for supporting the students and junior researchers attending the conference, chairs of neighboring statistics departments for encouraging their students to attend, and so on. Last but not least, we thank all the participants. Without your involvement, the event could not be a success. Go Gators!

Samuel Wu and Somnath Datta

Co-chairs, 2022 ICSA Applied Statistics Symposium Organizing Committee

Executives and Members of the Committees

Executives

President: Zhezhen Jin (zj7@cumc.columbia.edu)

Past President: Colin Wu (wuc@nhlbi.nih.gov)

President-elect: Gang Li (vli@ucla.edu)

Executive Director: Mengling Liu (executive.director@icsa.org)

ICSA Treasurer: Rui Feng (treasurer@icsa.org)

The ICSA Office Manager: Grace Ying Li, Email: office@icsa.org, Phone: (317) 287-4261

Board of Directors

Jason Liao (2020–2022, jliao@incyte.com)

Bin Nan (2020–2022, nanb@uci.edu)

Peihua Qiu (2020–2022, pqi@phhp.ufl.edu)

Jane Zhang (2020–2022, jane.zhang@abbvie.com)

Yichuan Zhao (2020–2022, yichuan@gsu.edu)

Shu-Hui Chang (2021–2023, shuhui@ntu.edu.tw)

Yong Chen (2021–2023, ychen123@mail.med.upenn.edu)

Chenlei Leng (2021–2023, c.leng@warwick.ac.uk)

Xiaodong Luo (2021–2023, xiaodong.luo@sanofi.com)

Yang Song (2021–2023, Yang_Song@vrtx.com)

Gang Li (2022–2024, Gang_Li@eisai.com)

Annie Qu (2022–2024, aqu2@uci.edu)

Lan Wang (2022–2024, lanwang@mbs.miami.edu)

Hao (Helen) Zhang (2022–2024, hzhang@math.arizona.edu)

Xingqiu Zhao (2022–2024, xingqiu.zhao@polyu.edu.hk)

Standing Committees

Program Committee

Chair: Hulin Wu (Hulin.Wu@uth.tmc.edu)

Members:

Aiyi Liu (2020–2022, JSM Representative 2021, liua@mail.nih.gov)

Pei Wang (2021–2023, JSM Representative 2022, pei.wang@mssm.edu)

Jianguo (Tony) Sun (2022–2024, JSM Representative 2023, sunj@missouri.edu)

Guoqing Diao (2020–2022, ICSA Symposium 2021, gdiao@email.gwu.edu)

Samuel Wu (2021–2023, ICSA Symposium 2022, samwu@biostat.ufl.edu)

Jian Kang (2022–2024, ICSA Symposium 2023, jiankang@umich.edu)

Gongjun Xu (2022–2024, ICSA Symposium 2023, gongjun@umich.edu)

Hongzhe Lee (2021–2022, ICSA International Conference 2019,
hongzhe@penntmedicine.upenn.edu)

Xin-Yuan Song (2021–2022, ICSA International Conference 2022, xysong@sta.cuhk.edu.hk)

Qingning Zhou (2020–2022, qzhou8@uncc.edu)

Liang Zhu (2020–2022, Liang.Zhu@eisai.com)

Jie Chen (2020–2022, jiechen0713@gmail.com)

Lihui Zhao (2022–2024, lihui.zhao@northwestern.edu)

Awards Committee

Chair: Chunming Zhang (2022, cmzhang@stat.wisc.edu)

Members:

Hongyuan Cao (2020–2022, hcao@fsu.edu)

Xiaofeng Shao (2020–2022, xshao@illinois.edu)

Xiaogang Su (2020–2022, xsu@utep.edu)

Yichao Wu (2020–2022, yichaowu@uic.edu)

Mingxiu Hu (2021–2023, mhu@nektar.com)

Jianxin Shi (2021–2023, jianxin.shi@nih.gov)

Ying Wei (2021–2023, yw2148@cumc.columbia.edu)

Jie Peng (2021–2023, jiepeng@ucdavis.edu)

Bo Huang (2022–2024, Bo.Huang@pfizer.com)

Charles Ma (2022–2024, tianzhou.ma0105@gmail.com)

Jiayang Sun (2022–2024, jsun@case.edu)

Nominating and Election Committee

Chair: Wenqing He (2022, whe@stats.uwo.ca)

Members:

Hailong Cheng (2020–2022, hailong.cheng@sunovion.com)

Bin Zhang (2020–2022, Bin.Zhang@cchmc.org)

Yuanyuan Lin (2022–2023, ylin@stat.cuhk.edu.hk)

Bo Fu (2022–2023, bo.fu.stat@gmail.com)

Sijian Wang (2022–2023, sijian.wang@stat.rutgers.edu)

Henry Horng-Shing Lu (2022–2023, hslu@stat.nycu.edu.tw)

Special Lecture Committee

Chair: Hongzhe Lee (2022, hongzhe@pennmedicine.upenn.edu)

Members:

Aiyi Liu (2021–2023, liua@mail.nih.gov)

Gang Li (2021–2023, Gang_Li@eisai.com)

Jianguo Sun (2022–2024, sunj@missouri.edu)

Xiaonan Xue (2022–2024, xiaonan.xue@einsteinmed.org)

Publication Committee

Chair: Linda Zhao (2022, lzhao@wharton.upenn.edu)

Members:

Hongkai Ji (2022–2024, Co-Editors of SIB, hji@jhu.edu)

Joan Hu (2021–2023, Co-Editors of SIB, joan_hu@sfu.ca)

Rong Chen (2021–2023, Co-Editors of Statistica Sinica, rongchen@stat.rutgers.edu)

Su-Yun Huang (2021–2023, Co-Editors of Statistica Sinica, syhuang@stat.sinica.edu.tw)

Xiaotong Shen (2021–2023, Co-Editors of Statistica Sinica, xshen@umn.edu)

Ding-Geng (Din) Chen (2020–2023, Editor of ICSA book series, dinchen@email.unc.edu)

Ming Wang (2021–2023, Editor for ICSA Bulletin, mwang@phs.psu.edu)

Mengling Liu (2020–2022, Executive Director of ICSA, executive.director@icsa.org)

Sheng Luo (2022–2024, sheng.luo@duke.edu)

Ying Ding (2022–2024, yingding@pitt.edu)

Membership Committee

Chair: Zhigen Zhao (2022, zhaozhg@temple.edu)

Members:

Shuwei Li (2020–2022, lishuwstat@163.com)

Yifei Sun (2021–2023, ys3072@cumc.columbia.edu)

Fei Huang (2022–2024, feihuang@unsw.edu.au)

Xun Chen (2022–2024, xun.chen@sanofi.com)

Wei Zhang (2022–2024, wei.zhang@boehringer-ingelheim.com)

Anru Zhang (2022–2024, anru.stat@gmail.com, anru.zhang@duke.edu)

IT Committee

Chair: Chengsheng Jiang (2021, 2022, website@icsa.org)

Archive Committee

Chair: Naitee Ting (2022, naitee.ting@boehringer-ingelheim.com)

Members:

Xin (Henry) Zhang (2021–2023, henry@stat.fsu.edu)

Rui Miao (2021–2023, ruimiao@gwu.edu)

Xin Tian (2022–2024, tianx@nhlbi.nih.gov)

Jun Yan (2022–2024, jun.yan@uconn.edu)

Finance Committee

Chair: Rui Feng (2022–2024, ruifeng@penncmedicine.upenn.edu)

Members:

Xin He (2021–2023, xinhe@umd.edu)

Rochelle Fu (2022–2024, fu@ohsu.edu)

Financial Advisory Committee

Chair: Fang Chen (FangK.Chen@sas.com)

Members:

Hongliang Shi (2020–2022, hongliangshi15@gmail.com)

Nianjun Liu (2021–2023, liunian@indiana.edu)

Xiangqin Cui (2021–2023, xiangqin.cui@emory.edu)

Rochelle Fu (2022–2024, fu@ohsu.edu)

Yuan Jiang (2022–2024, yuan.jiang@stat.oregonstate.edu)

Lingzi Lu Award Committee (ASA/ICSA)

Chair: Ivan Chan (2022, ivan.chan@bms.com)

Members:

Shelly Hurwitz (2017–2022, hurwitz@hms.harvard.edu)

Laura J Meyerson (2020–2022, laurameyerson@msn.com)

Kelly Zou (2022–2024, kelly.zou@viatris.com)

ICSA Representative to JSM Program Committee

Pei Wang (2022, pei.wang@mssm.edu)

Jianguo (Tony) Sun (2023, sunj@missouri.edu)

Ad hoc Committees

2022 Applied Statistics Symposium

Chair: Samuel Wu (samwu@biostat.ufl.edu)

2022 JSM Local Committee

Chair: Guoqing Diao (gdiao@gwu.edu)

2022 ICSA China Conference

Chair: Yingying Fan (fanyingy@usc.edu)

2023 Applied Statistics Symposium

Chairs: Jian Kang (jiankang@umich.edu) and Gongjun Xu (gongjun@umich.edu)

Chapters

ICSA-Canada Chapter

President: Yingwei Paul Peng, Queen's University (2021–2022)

Past-President: Liqun Wang, University of Manitoba (2021–2022)

President-Elect: Joan Hu, Simon Fraser University (2021–2022)

Secretary/Treasurer: Leilei Zeng, University of Waterloo (2019–2024)

ICSA-Midwest Chapter

Chair: Xiaohong Huang (midwest@icsa.org, xiaohong.huang@abbvie.com), Abbvie Inc.

ICSA-Taiwan Chapter

Chair: Chao A. Hsiung (hsiung@nhri.org.tw)

Executive Committee

Chair: Peihua Qiu (University of Florida)

Members:

Somnath Datta (University of Florida)

Zhezhen Jin (Columbia University)

Ji-Hyun Lee (University of Florida)

Mengling Liu (NYU Langone Health)

Tony Sun (University of Missouri)

Samuel Wu (University of Florida)

Scientific Program Committee

Chairs: Samuel Wu (University of Florida) and Somnath Datta (University of Florida)

Members:

Dipankar Bandyopadhyay (Virginia Commonwealth University)

Jie Chen (Augusta University)

Xinping Cui (UC Riverside)

Susmita Datta (University of Florida)

Adam Ding (Northeastern University)

Joan Hu (Simon Frazer University)

Jianhua Huang (CUHK)

Victor Hugo Lachos Davila (University of Connecticut)

Hongzhe Lee (University of Pennsylvania)

Gang Li (UCLA)

Tapabrata Maiti (Michigan State University)

Rajeshwari Sundaram (NIH/NICHHD)

Ming Wang (Pennsylvania State University)

Wei Wu (Florida State University)

Dong Xi (Gilead Sciences)
Gongjun Xu (University of Michigan)
Hongyu Zhao (Yale University)
Jun Zhao (Astellas)
Yichuan Zhao (Georgia State University)
Kelly Zou (Viatris)

Local Committee

Chair: Ji-Hyun Lee (University of Florida)

Members:

Kristen Cason (University of Florida)
Melissa Layne (University of Florida)

Student Paper Competition Committee

Chair: Zhigang Li (University of Florida)

Members:

Hsin-wen Chang (Institute of Statistical Science, Academia Sinica)
Guanhua Chen (University of Wisconsin-Madison)
Li Chen (Indiana University)
Subharup Guha (University of Florida)
Pengsheng Ji (University of Georgia)
Fan Li (Yale University)
Muxuan Liang (University of Florida)
Xiangyang Lou (University of Florida)
Qing Lu (University of Florida)
Jing Ma (Fred Hutchinson Cancer Research Center)
Guogen Shan (University of Florida)

Zihua Su (University of Florida)
Zhengzheng Tang (University of Wisconsin-Madison)
Xuefeng Wang (Moffitt Cancer Center)
Feifei Xiao (University of Florida)
Kai Yang (Medical College of Wisconsin)
Yang Yang (University of Florida)
Lu You (University of South Florida)
Lihui Zhao (Northwestern University)

Short Course Committee

Chair: Susmita Datta (University of Florida)

Poster Committee

Chair: Guogen Shan (University of Florida)

Member:

Rhonda Bacher (University of Florida)
Li Chen (Indiana University)
Jonathan Fischer (University of Florida)
Steven Foti (University of Florida)
Matt Hitchings (University of Florida)
Zhigang Li (University of Florida)
Muxuan Liang (University of Florida)
Ruitao Lin (MD Anderson Cancer Center)
Qing Lu (University of Florida)
Arlene Naranjo (University of Florida)
Robert Parker (University of Florida)
Qinglin Pei (University of Florida)

Arkaprava Roy (University of Florida)

Wei Xue (University of Florida)

Fundraising Committee

Chair: Jane Zhang (AbbVie)

Members:

Jessica Cai (Regeneron)

Pranab Mitra (Takeda)

Furong Sun (Regeneron)

Daisy Yuan (AbbVie)

Financial/Business Committee

Chair: Renee Douglass (University of Florida)

Members:

Rochelle Fu (ICSA Treasurer)

Deana Nance (University of Florida)

Guogen Shan (University of Florida)

Program Book and Website Committee

Chairs: Xiangyang Lou (University of Florida) and Qinglin Pei (University of Florida)

Members:

Yunfeng Dai (University of Florida)

Hanzhi Gao (University of Florida)

Chengsheng Jiang (ICSA IT)

Guanhong Miao (University of Florida)

Sponsors Information

The 2022 ICSA Applied Statistics Symposium Program Committees gratefully appreciate National Science Foundation for supporting the individuals below.



University of Florida:

Animesh Mitra	Archie Sachdeva	Chang Jiang	David Lindberg
Dongyuan Wu	Dorothy Ellis	Eleni Dilma	Fan Yi
Guangjin Luo	Guanhong miao	Hanzhi Gao	Heejun Shin
Hongqiang Sun	Jiayuan Zhou	Lei Yang	Li Duan
Luer Zhong	Man Chong Leong	Matthew Corcoran	Meilin Jiang
Natalie DelRocco	Nilanjana Chakraborty	Peter Chang	Rongzi Liu
Samuel Anyaso-Samuel	Sara Nutley	Saurabh Bhandari	Seungjun Ahn
Shangchen Song	Shoumi Sarkar	Srijata Samanta	Tingting Hou
Vincent Mei	Wangze Yu	Xiaoru Dong	Xiaoxi Zhang
Xinyu Yan	Xiulin Xie	Yeison Quiceno Duran	Yi Xu
Yipeng Wang	Yuan Zhou	Yue Wu	Yutao Zhang
Yuting Yang	Zibo Tian		

University of Central Florida:

Hayden Hampton	Imani Eatman	Ngoc Ty Nguyen	Qing He
Trymain Rivero	Victor Nkwocha	Yi Liu	

Florida State University:

Anisha Das	Avizit Adhikary	Durbadal Ghosh	Sudipto Saha
Sunghee Park			

Other universities:

Andrew Chen	Dongyue Xie	Doudou Zhou	Jeffrey Wu
Jiuzhou Wang	Kan Chen	Kelin Zhong	Lin Ge
Linquan Ma	Lu Zhang	Margaret Banker	Maxine Yu
Menglu Liang	Samaneh Nasiri	Samrat Roy	Shuting Shen
Xiaoqing Tian	Xiaotian Zheng	Xinkai Zhou	Yangjianchen Xu
Ying Jin			

Sponsors Information

The 2022 ICSA Applied Statistics Symposium Program Committees gratefully acknowledge the generous support of our sponsors below.

Gold sponsors



Sponsors Information

Silver sponsors



Bronze sponsors



abbvie

PEOPLE. PASSION. POSSIBILITIES.®

AbbVie is a global, research-driven biopharmaceutical company committed to developing innovative advanced therapies for some of the world's most complex and critical conditions. The company's mission is to use its expertise, dedicated people and unique approach to innovation to markedly improve treatments across four primary therapeutic areas: immunology, oncology, virology and neuroscience. In more than 75 countries, AbbVie employees are working every day to advance health solutions for people around the world. For more information about AbbVie, please visit us at www.abbvie.com. Follow @abbvie on Twitter, Facebook or LinkedIn.



**CANCER HAS
NO BORDERS.
NEITHER
DO WE®**

**A global biotech innovator focused
on improving treatment
outcomes and patient access**

BeiGene is a global, science-driven biotechnology company focused on developing innovative and affordable medicines to improve treatment outcomes and access for patients worldwide. With a broad portfolio of more than 40 clinical candidates, we are expediting development of our diverse pipeline of novel therapeutics through our own capabilities and collaborations. We are committed to radically improving access to medicines for two billion more people by 2030. BeiGene has a growing global team of over 6,900 colleagues across five continents.

To learn more about BeiGene,
please visit www.beigene.com and follow us
on Twitter at [@BeiGeneGlobal](https://twitter.com/BeiGeneGlobal).

 www.linkedin.com/company/beigene

 [@BeiGeneGlobal](https://twitter.com/BeiGeneGlobal)

BeiGene

Value through
innovation

*Improving the health
of humans and animals
– Our goal.*

Family-owned since 1885, Boehringer Ingelheim is one of the leading pharmaceutical companies worldwide. More than 51,000 employees create value through innovation in the business areas Human Pharma, Animal Health and Biopharmaceutical Contract Manufacturing. In our role as our patients' partner we concentrate on researching and developing innovative medicines and therapies that can improve and extend patients' lives.

www.boehringer-ingelheim.com





Transforming patients' lives through science™

At Bristol Myers Squibb, we are inspired by a single vision - transforming patients' lives through science.

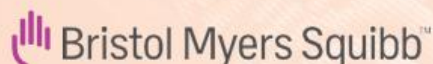
We are in the business of breakthroughs - the kind that transform patients' lives through life-saving, innovative medicines. We have the most talented people in the industry who come to work every day dedicated to our mission of discovering, developing and delivering innovative medicines that help patients prevail over serious diseases.

We combine the agility of a biotech with the reach and resources of an established pharmaceutical company to create a global leading biopharma company.

In oncology, hematology, immunology and cardiovascular disease - and one of the most diverse and promising pipelines in the industry - we focus on innovations that drive meaningful change. We bring a human touch to every treatment we pioneer. With great pride, we celebrate each time our patients take back their lives.

Our shared values are central to who we are, what we do and how we do it. Passion, innovation, urgency, accountability, inclusion and integrity ground our work and unite our community. We never give up in our search for the next innovation that could mean new hope for patients who are urgently seeking new treatment options today.

Our symbol, the hand, is a simple, universal expression of healing, of giving and receiving care. It is a representation of humanity, of the personal touch we bring to our work and to every treatment we pioneer. Our brand fully embodies our vision, and embraces our commitment to compassionate science and putting patients and people first.



Visit [bms.com](https://www.bms.com) to see how we're bringing a human touch to everything we do.

ClinChoice

The Standard of Excellence

• 25+ Year History of Quality • Flexible Approach • Expedited Timelines • Global Resourcing

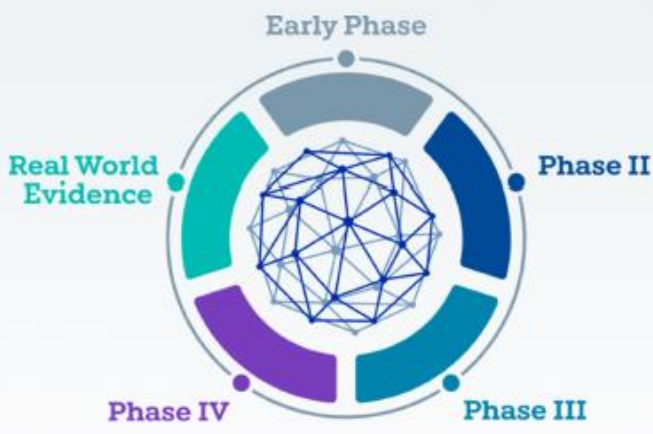


Your Full Service Clinical CRO Partner

Pharma | Biotech | Med Device |
Cosmetics | Consumer Healthcare

Services for the Full Development Lifecycle

Accelerating drug and device approvals to market with post-market support for more than 25 years



4 Year History As An ICSA Sponsor!

www.clinchoice.com
Formerly FMD K&L

Regeneron is a leading biotechnology company that invents life-transforming medicines for people with serious diseases.

Regeneron (NASDAQ: REGN) is a leading biotechnology company that invents life-transforming medicines for people with serious diseases. Founded and led for nearly 35 years by physician-scientists, our unique ability to repeatedly and consistently translate science into medicine has led to nine FDA-approved treatments and numerous product candidates in development, almost all of which were homegrown in our laboratories. Our medicines and pipeline are designed to help patients with eye diseases, allergic and inflammatory diseases, cancer, cardiovascular and metabolic diseases, pain, hematologic conditions, infectious diseases and rare diseases.

REGENERON
SCIENCE TO MEDICINE®

Regeneron is accelerating and improving the traditional drug development process through our proprietary *VelociSuite*® technologies, such as *VelocImmune*®, which uses unique genetically humanized mice to produce optimized fully human antibodies and bispecific antibodies, and through ambitious research initiatives such as the Regeneron Genetics Center®, which is conducting one of the largest genetics sequencing efforts in the world. For additional information about the company, please visit www.regeneron.com or follow @Regeneron on Twitter.

General Company Information

- Founded in 1988: Publicly traded company (NASDAQ: REGN) since 1991
- More than 10,000 employees in the U.S., Canada, UK and EU
- 2021 R&D investment of \$2.9 billion



Locations

- Tarrytown, NY: Corporate and Research & Development headquarters
- Rensselaer, NY and Limerick, Ireland: Large-scale biologics Industrial Operations and Product Supply (IOPS) facilities
- Sleepy Hollow, NY, Basking Ridge, NJ, and Washington, D.C.: offices
- Amsterdam, Dublin, London, Munich and Toronto: Global business offices



Leadership Team

- **Leonard S. Schleifer, MD, PhD**
President and Chief Executive Officer
+ Fellow, American Association for the Advancement of Science (AAAS)
- **George D. Yancopoulos, MD, PhD**
President and Chief Scientific Officer
+ Member, National Academy of Sciences
- **P. Roy Vagelos, MD**
Chairman of the Board
+ Former Chief Executive Officer and Chairman of the Board, Merck & Co.
+ Member, National Academy of Sciences
- **Board of Directors** includes two Nobel Laureates and seven members of the National Academy of Sciences



FDA-Approved & Marketed Medicines*



① In collaboration with Sanofi outside of U.S. For Praluent, in collaboration with Sanofi prior to April 2020; effective April 2020, Regeneron is solely responsible for the U.S. development and commercialization and Sanofi is solely responsible for the ex-U.S. development and commercialization of Praluent.
② In collaboration with Bayer outside of U.S. | ③ Marketed by Sanofi. | ④ Marketed by Kiniksa Pharmaceuticals.

* U.S. Food and Drug Administration



Why Choose The Lotus Group?

The Lotus Group LLC is a certified minority women-owned (MWBE) business delivering biometrics recruiting services nationally. In 2020, we were recognized as one of the 50 fastest growing women-owned businesses by the Women Presidents' Organization, sponsored by American Express.

Our team members are located throughout the country including California, Florida, Massachusetts, and New Jersey; many with over 20 years of staffing experience in the pharmaceutical industry.

Our consultative approach and knowledge of the market ensure that your career is in good hands. We provide more than just access to great companies and industry-leading opportunities. We see ourselves as your "career advocates" as we provide you with everything you need to successfully navigate your job search and advance your career.


Step on to The Lotus Group bridge – connecting great biometrics candidates and companies – and start your journey to success!

Key Functional Areas We Support

- Biostatistics
- Clinical & Statistical Programming
- Data Management
- Data Science
- Epidemiology
- Pharmacometrics
- HEOR / RWE

- Check out our website: <https://tlqcareers.com/>
- Follow us on LinkedIn - scan the QR Code
- Send resumes or inquiries to info@tlqcareers.com





AMGEN FACT SHEET | 2022

About Amgen

Our Mission:
To Serve Patients

Amgen is committed to unlocking the potential of biology for patients suffering from serious illnesses by discovering, developing, manufacturing and delivering innovative human therapeutics. This approach begins by using tools like advanced human genetics to unravel the complexities of disease and understand the fundamentals of human biology.

Our belief—and the core of our strategy—is that innovative, highly differentiated medicines that provide large clinical benefits in addressing serious diseases are medicines that will not only help patients, but also will help reduce the social and economic burden of disease in society today.

Amgen focuses on areas of high unmet medical need and leverages its expertise to strive for solutions that improve health outcomes and dramatically improve people's lives. A biotechnology innovator since 1980, Amgen has grown to be one of the world's leading independent biotechnology companies, has reached millions of patients around the world and is developing a pipeline of medicines with breakaway potential.

Center for Design & Analysis

Our Mission:
To discover, develop & deliver life changing medicines to patients everywhere



A full service CRO built from a statistical and data management center of excellence.

We serve some of the best-known pharmaceutical, biotechnology, and medical device companies worldwide and help drive research for the most advanced drugs, biologics, and medical devices in development today. Our foundation is built on excellence in biometrics services and we have grown to become a data- and information-driven, quality and customer service-focused full service Contract Research Organization (CRO).

We are proud of our reputation as a high quality, customer-focused, and flexible clinical research partner for our rapidly expanding base of clients.

We're hiring! Accelerate your career at Everest.

- Biostatisticians (All Levels)
- Data Managers
- Statistical Programmers (All Levels)
- Project Managers
- Bilingual (English and Chinese) Medical Writers



- Over 350 Global Staff
- 18+ Years of Industry Excellence
- 30+ Therapeutic Areas
- 80+ Active Clients
- 30%+ Projects in Oncology

CREATING POSSIBLE

Gilead is proud to support the International Chinese Statistical Association 2022 Applied Statistics Symposium.



For more information,
please visit www.gilead.com



For more than 125 years, Merck, known as MSD outside of the United States and Canada, has been inventing for life, bringing forward medicines and vaccines for many of the world's most challenging diseases in pursuit of our mission to save and improve lives. We demonstrate our commitment to patients and population health by increasing access to health care through far-reaching policies, programs and partnerships. Today, Merck continues to be at the forefront of research to prevent and treat diseases that threaten people and animals — including cancer, infectious diseases such as HIV and Ebola, and emerging animal diseases — as we aspire to be the premier research-intensive biopharmaceutical company in the world. For more information, visit www.merck.com and connect with us on Twitter, Facebook, Instagram, YouTube and LinkedIn.



We are an innovative global healthcare company, driven by one purpose: we chase the miracles of science to improve people's lives.

We are dedicated to transforming the practice of medicine by working to turn the impossible into the possible. We provide potentially life-changing treatment options and life-saving vaccine protection to millions of people globally, while putting sustainability and social responsibility at the center of our ambitions.

Our Biostatistics and Programming department plays a key role by investing in the development of all our team members, offering compelling and exciting career opportunities that value diversity of thought and abilities. We have a shared commitment to bring innovation and rigor to our competitive portfolio, including in oncology immunology, neurology and one of the largest rare disease pipelines in the industry.

To learn more about Sanofi, please visit www.sanofi.com and also see <https://www.sanofi.us/en/about-us/awards-and-recognitions>

sanofi

Follow us     

Statistical & Quantitative Sciences

VISION
Better health and a brighter future through innovative statistical and data sciences

MISSION
Bring transformative therapies to patients by harnessing the power of data with:

- Innovative trial design and development strategies
- Advanced statistical and data science methods & modeling
- Best-in-class data and programming engine driven by automation

Takeda is embracing data, digital and technology across its R&D efforts. Through the Data Sciences Institute (DSI), we are powering our development engine and enabling next generation clinical trial design and insights to bring our potential medicines to patients faster. Come be a

Sponsors Information

part of this exciting vision and join our growing team. Check out our career opportunities here: <https://www.takedajobs.com/research-and-development>.



Vertex is among the most innovative companies in the pharmaceutical industry, having discovered and developed the only approved therapies for cystic fibrosis, treating the underlying cause of the disease, and is now expanding to multiple therapeutic areas, such as sickle cell disease and Duchenne muscular dystrophy through our new Vertex Cell and Genetic Therapies (VCGT) group. As a medium-sized biotech, we offer the solid training and experience of the large pharmaceuticals, while retaining the speed and agility of the small biotechs.

Positions are available based at our Boston Seaport Headquarters. Please contact Vanessa Marin at vanessa_marin@vrtx.com



Servier Pharmaceuticals LLC is a commercial-stage company with a passion for innovation and improving the lives of patients, their families and caregivers. A privately held company, Servier has the unique freedom to devote its time and energy toward putting those who require our treatment and care first, with future growth driven by innovation in areas of unmet medical need.

As a growing leader in oncology, Servier is committed to finding solutions that will address today's challenges. The company's oncology portfolio of innovative medicines is designed to bring more life-saving treatments to a greater number of patients, across the entire spectrum of disease and in a variety of tumor types.

Servier believes co-creation is fundamental to driving innovation and is actively building alliances, acquisitions, licensing deals and partnerships that bring solutions and accelerate access to therapies. With our commercial expertise, global reach, scientific expertise and commitment to clinical excellence, Servier Pharmaceuticals is dedicated to bringing the promise of tomorrow to the patients that we serve.

Accommodation

The ICSA 2022 Applied Statistics Symposium offers accommodations for conference guests at [Hotel Eleo](#), [Aloft](#), and [Hilton University of Florida conference center](#).

Shuttle service will be provided between the hotels and the venue. In addition, daily parking will be available upon request.

Venue of the Symposium

The ICSA 2022 Applied Statistics Symposium is located at the **HPNP (Health Professions, Nursing & Pharmacy) Building** on Center Drive, just north of the University of Florida Medical Center/Shands Hospital (i.e., 1225 Center Dr, Gainesville, FL 32611).

ICSA Banquet Map

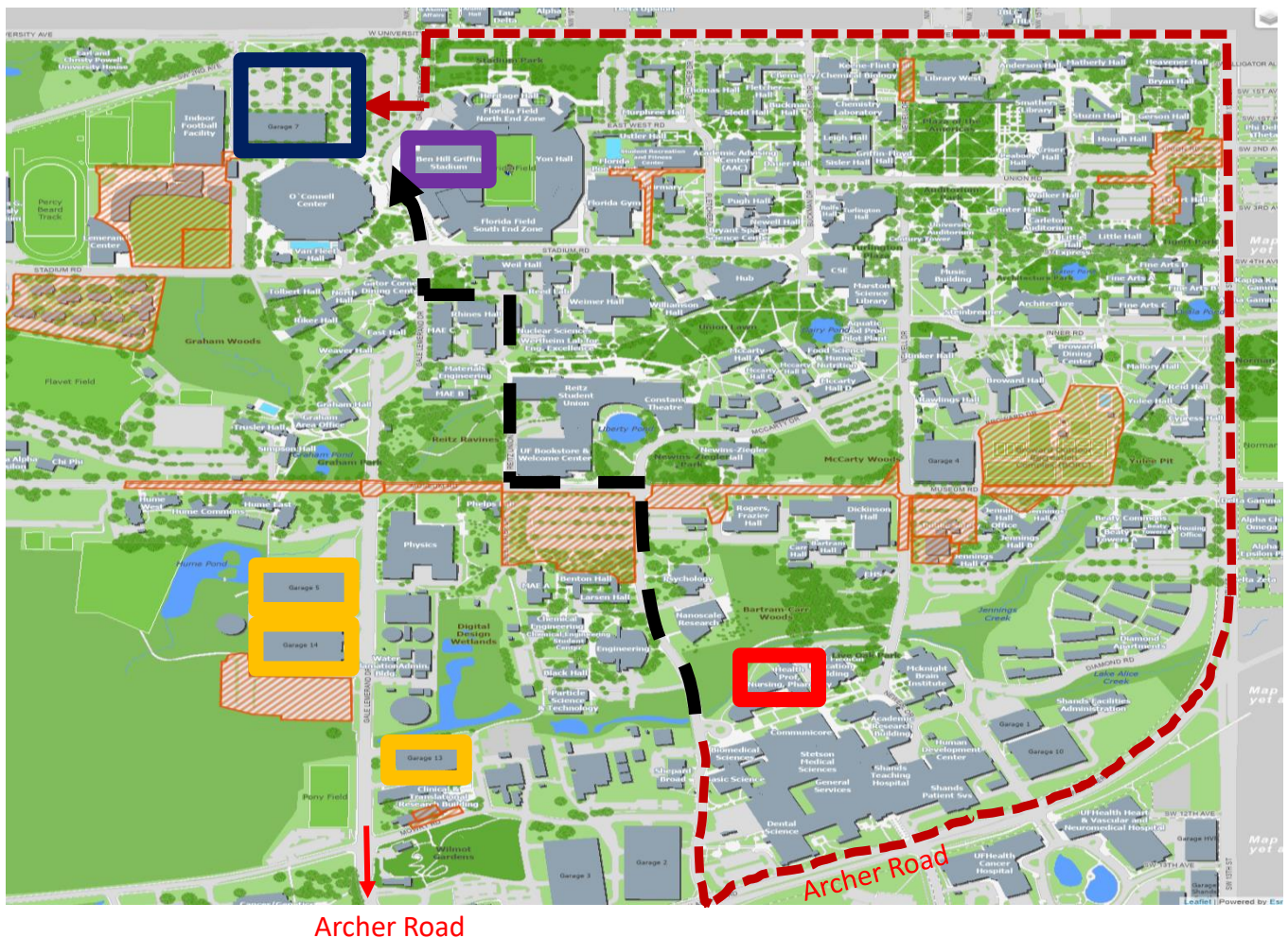
Follow directions below. Upon arrival at the Ben Hill Griffin Stadium look for signage and event staff. Enter the Champions Club beside the Heisman Statues near the Presidential Suite. Take elevators to the 5th floor.



Walking direction

Driving direction

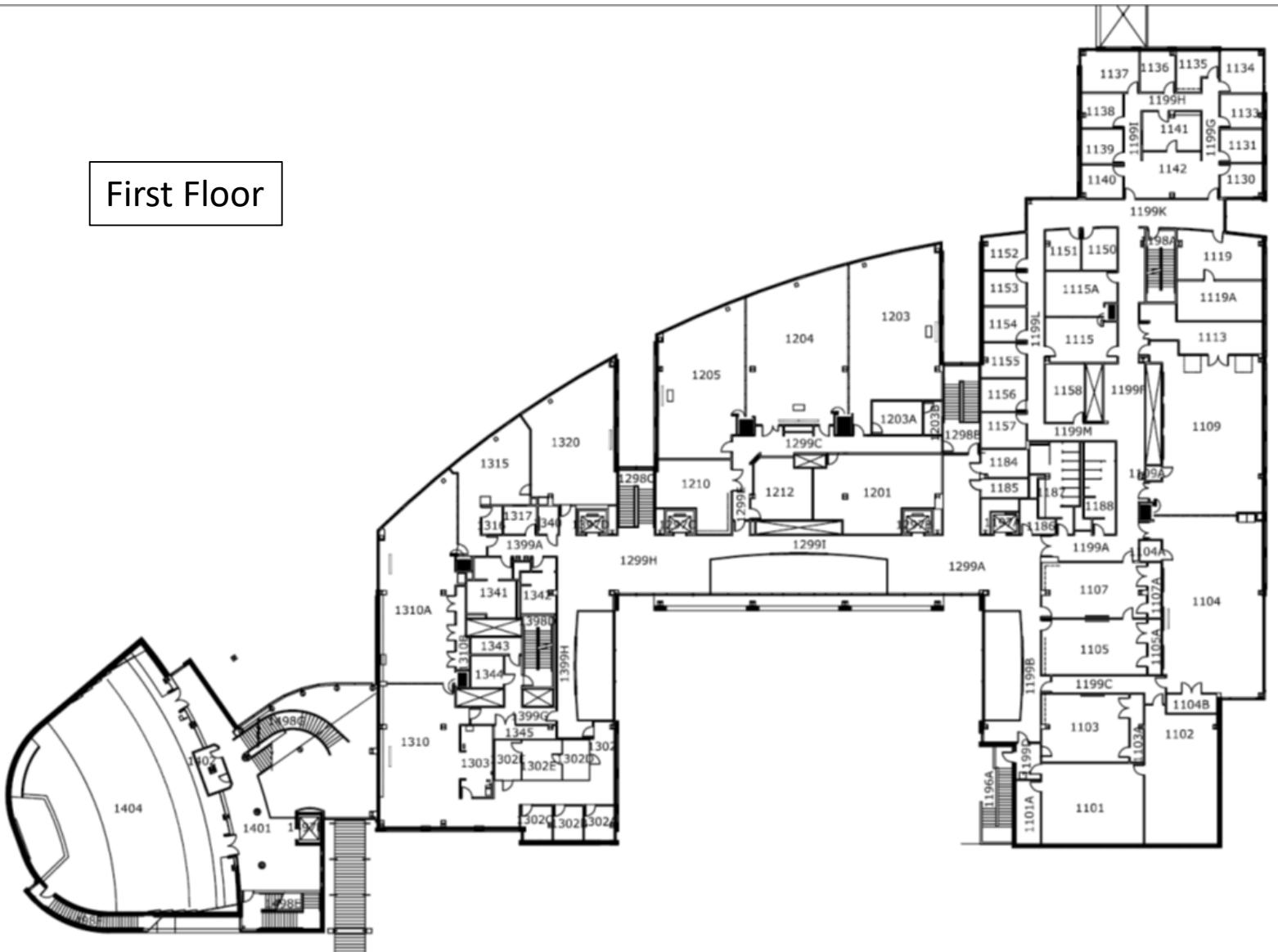
1. Head south from campus and tune left (east) on SW Archer Road
2. Left (north) on SW 13th Street
3. Left (west) on University Avenue
4. Left (south) on Gale Lemerand Rd.
5. Take 1st right into the O'Connell Center parking lot.





Session	Room
Plenary	HPNP 1404
A	HPNP G312
B	HPNP G112
C	HPNP G101
D	HPNP G103
E	HPNP G114
F	HPNP G301
G	HPNP 1101
H	HPNP 1102

First Floor



LOCAL POINTS OF INTEREST

Fine Dining

Blue Gill Quality Food 1310 SW 13th St., Gainesville 352-872-5181

<https://bluegillqualityfoods.com/>

Chopstix Cafe 3500 SW 13th St., Gainesville 352-367-0003

<https://chopstixcafe.business.site/>

Crane Ramen 16 SW 1st Ave., Gainesville 352-727-7422

<https://craneramen.com/>

Dragonfly Sushi 201 SE 2nd Ave., Gainesville 352-371-3359

<http://www.dragonflyrestaurants.com/gainesville-florida/>

Embers Wood Grill 3545 SW 34th St., Gainesville 352-380-0901

<https://embersofflorida.com/>

Sensei Asian Bistro 4860 SW 31st Pl., Gainesville 352-214-1239

<https://www.insensei.com/>

Spurrier's Gridiron Grille 4860 SW 31st Pl., Gainesville 352-500-4422

<https://spurriers.com/>

Sohao Cafe 3045 SW 34th St., Gainesville 352-792-6789

<http://www.sohaocafe.com/>

Shopping

Celebration Pointe 4949 Celebration Pointe Ave., Gainesville

352-204-9008 <https://www.celebrationpointe.com/>

Haile Village Center 5100 SW 91st Ter., Gainesville 352-378-9793

<http://www.hailevillagecenter.com/default.php>

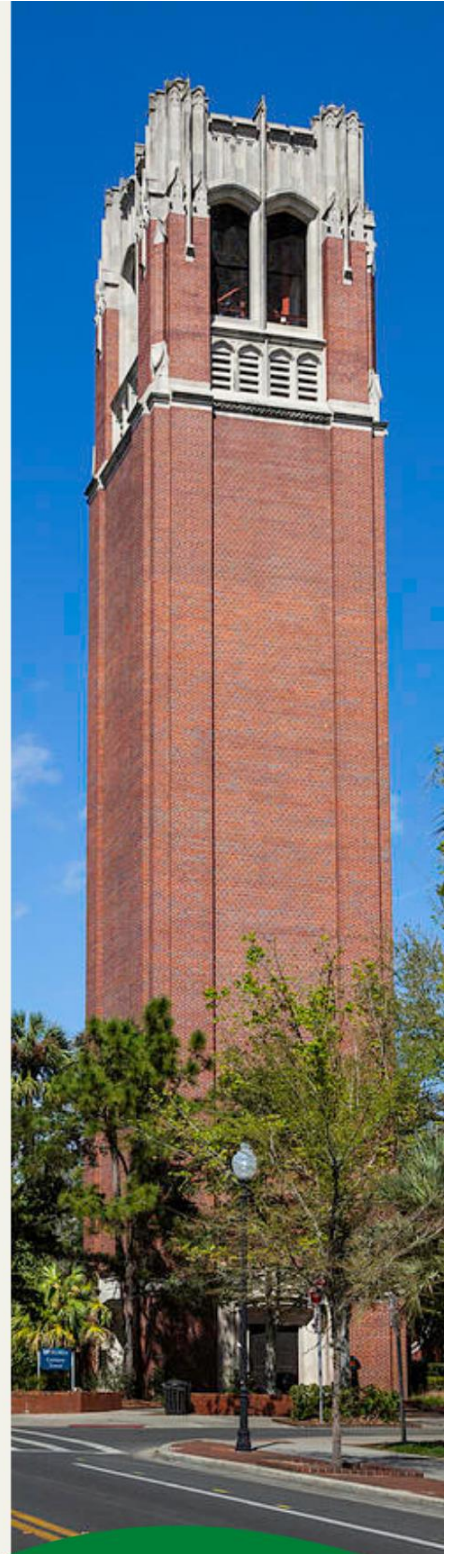
The Oaks Mall 6419 W Newberry Rd., Gainesville 352-331-0040

<https://www.theoaksmall.com/en.html>

Thornebrook Village 2441 NW 43rd St., Gainesville 352-378-4947

<https://thornebrookvillage.com/>

- **St. Johns Town Center** 4663 River City Dr., Jacksonville
904-998-7156 <https://www.simon.com/mall/st-johns-town-center>
- **Orlando Outlet Marketplace** 5269 International Dr., Orlando
407-352-9600 <https://www.premiumoutlets.com/outlet/orlando-outlet-marketplace>
- **St. Augustine Premium Outlets** 2700 FL-16, St. Augustine
904-825-1555 <https://www.premiumoutlets.com/outlet/st-augustine>



VISIT GAINESVILLE

33 North Main St., Gainesville FL 32601

www.visitgainesville.com

Nature

Butterfly Rainforest 3215 Hull Rd., Gainesville 352-846-2000

<https://www.floridamuseum.ufl.edu/exhibits/butterfly-rainforest/>

Depot Park 874 SE 4th St., Gainesville 352-393-8510

<http://www.depotpark.org/>

Kanapaha Botanical Gardens 4700 SW 58th Dr., Gainesville 352-372-4981 <https://kanapaha.org/>

Sweetwater Wetland Park 325 SW Williston Rd., Gainesville 352-393-8520 <https://www.sweetwaterwetlands.org/>

UF Bat Houses Museum Road across from Lake Alice 352-846-2000 <https://www.floridamuseum.ufl.edu/bats/>

- **Gilchrist Blue Springs State Park**

7450 NE 60th St., High Springs 386-454-1369

<https://www.floridastateparks.org/parks-and-trails/ruth-b-kirby-gilchrist-blue-springs-state-park>

Museums

Cade Museum for Creativity & Invention

811 S Main St., Gainesville 352-371-8001

<https://www.cademuseum.org/>

Harn Museum of Art 3529 Hull Rd., Gainesville

352-392-9826 <https://harn.ufl.edu/>

Breweries and Wineries

Bluefield Estate Winery 22 NE CR 234, Gainesville

352-337-2544 <https://www.facebook.com/BluefieldEstateWinery/>

Cypress and Grove Brewing Company 1001 NW 4 St., Gainesville

352-376-4993 <https://www.cypressandgrove.com/>

First Magnitude Brewing Company 1220 SE Veitch St., Gainesville

352-727-4677 <https://fmbrewing.com/>

Swamp Head Brewery 3650 SW 42nd Ave., Gainesville

352-505-3035 <https://swamphead.com/>



Brought to you by:



Dining and WiFi Information

To access UF WI-FI:

First, join the "ufguest" wireless network from your mobile device. When you attempt to browse to the Internet, your browser will be redirected to a portal web page. This web page presents an Acceptable Use Policy (AUP) with a button at the bottom to indicate you accept the policy. Once the policy is read and accepted, you will be permitted to access the Internet.

Program Overview

ICSA 2022 Applied Statistics Symposium (Program and Activity Overview)

Theme: Statistical Innovations in the Era of Artificial Intelligence and Data Science

Date	Time	Event	Location																				
Sunday, June 19	7:30-18:30	Registration	HPNP Lobby																				
	8:30-12:30	Short Courses SC01, SC02, SC03, SC04	See Program Book																				
	12:30-13:30	Lunch																					
	13:30-17:30	Short Courses SC02, SC03, SC05, SC06	See Program Book																				
	18:30-22:00	Welcome Reception and Mixer	Hilton Hotel																				
Monday, June 20	7:30-18:30	Registration	<table border="1"> <thead> <tr> <th>Session</th> <th>HPNP Room</th> </tr> </thead> <tbody> <tr> <td>Plenary</td> <td>1404</td> </tr> <tr> <td>A</td> <td>G312</td> </tr> <tr> <td>B</td> <td>G112</td> </tr> <tr> <td>C</td> <td>G101</td> </tr> <tr> <td>D</td> <td>G103</td> </tr> <tr> <td>E</td> <td>G114</td> </tr> <tr> <td>F</td> <td>G301</td> </tr> <tr> <td>G</td> <td>1101</td> </tr> <tr> <td>H</td> <td>1102</td> </tr> </tbody> </table>	Session	HPNP Room	Plenary	1404	A	G312	B	G112	C	G101	D	G103	E	G114	F	G301	G	1101	H	1102
	Session	HPNP Room																					
	Plenary	1404																					
	A	G312																					
	B	G112																					
	C	G101																					
	D	G103																					
	E	G114																					
	F	G301																					
	G	1101																					
	H	1102																					
8:00-8:30	Welcome and Opening Remarks																						
8:30-9:30	Plenary Keynote Talk 1																						
9:30-10:00	Coffee Break																						
10:00-11:40	Invited Sessions 1A to 1H																						
11:40-13:00	Lunch																						
13:00-14:40	Invited Sessions 2B to 2H																						
14:34-15:00	Coffee Break																						
15:00-16:40	Invited Sessions 3A to 3H																						
16:40-17:00	Coffee Break																						
17:00-18:40	Invited Sessions 4A to 4H																						
19:00-20:00	Poster Session																						
Tuesday, June 21	7:30-18:30	Registration																					
	8:30-9:30	Plenary Keynote Talk 2																					
	9:30-10:00	Coffee Break																					
	10:00-11:40	Invited Sessions 5A to 5H																					
	11:40-13:00	Lunch																					
	13:00-14:40	Special invited Session																					
	14:40-15:00	Coffee Break																					
	15:00-16:40	Invited Sessions 6A to 6H																					
	16:40-17:00	Coffee Break																					
	17:00-18:40	Invited Sessions 7A to 7H																					
	19:00-22:00	Conference Banquet																					
Wednesday, June 22	8:30-9:30	Plenary Keynote Talk 3																					
	9:30- 10:00	Coffee Break																					
	10:00-11:40	Invited Sessions 8A to 8H																					
	11:40-13:00	Lunch																					
	13:00-14:40	Invited Sessions 9A to 9H																					
14:40	Adjournment																						

Keynote Speaker



David O. Siegmund, Ph.D., who holds the John D. and Sigrid Banks Chair at Stanford University, Stanford, CA, is a statistician who is comfortable in both the airy heights of theory and the practicalities of real-world applications. He works at the interface between probability and statistics, applying the tools he develops to topics as diverse as the design of medical clinical trials and mapping the locations of genes that are involved in specific physiological traits. His work has earned him several awards, including a Guggenheim Fellowship in 1974, the Humboldt Prize in 1980, and membership in the American Academy of Arts and Sciences in 1994. In 2002 he was elected to the National Academy of Sciences. His Inaugural Article, published in this issue of PNAS, reviews recent methodological developments in quantitative trait locus mapping and addresses the problem of mapping with selected, rather than random, samples.

Location and Time: HPNP Auditorium (1404), June 20 (Monday), 8:30 am – 9:30 am

Organizer: ICSA special lecture committee

Keynote Host: Samuel Wu, Ph.D., University of Florida

Title: Change detection, estimation, and segmentation

Abstract: I will first discuss the maximum score statistic to detect and estimate via confidence regions change-points in the level, slope, or other property of a Gaussian process and to segment the process when there appear to be multiple changes. Sequential detection is also considered. Examples involving temperature variations, levels of atmospheric greenhouse gases, temporal incidence of hate crimes, suicide rates, incidence of Covid-19, and excess deaths during the Covid-19 pandemic illustrate the general theory.

I will describe research in progress for spatio-temporal processes, where the spatial features can be either (A) unstructured vectors of observations or (B) random fields where changes of interest are geometrically clustered. Examples include low and (perhaps sparse) high dimensional cases.

I also mention the special problems posed by temporal and/or spatial dependence. Failure to account for correlations can lead to inflated false positive rates. while the change-points themselves can lead to upwardly biased estimates of correlations that result in loss of power.

Aspects of this research involve collaboration with Fang Xiao, Li Jian, Liu Yi, Nancy Zhang, Benjamin Yakir and Li (Charlie) Xia.

Keynote Speaker



Jianqing Fan, Ph.D., is a statistician, financial econometrician, and data scientist. He is Frederick L. Moore'18 Professor of Finance, Professor of Statistics, and Professor of Operations Research and Financial Engineering at the Princeton University where he chaired the department from 2012 to 2015. He is the winner of The 2000 COPSS Presidents' Award, Morningside Gold Medal for Applied Mathematics (2007), Guggenheim Fellow (2009), Pao-Lu Hsu Prize (2013) and Guy Medal in Silver (2014).

Location and Time: HPNP Auditorium (1404), June 21 (Tuesday), 8:30 am – 9:30 am

Organizer: ICSA special lecture committee

Keynote Host: Somnath Datta, Ph.D., University of Florida

Title: Measuring housing activeness from multi-source big data and machine learning

Abstract: Measuring timely high-resolution socioeconomic outcomes is critical for policy-making and evaluation, but hard to reliably obtain. With the help of machine learning and cheaply available data such as social media and nightlight, it is now possible to predict such indices in fine granularity. This paper demonstrates an adaptive way to measure the time trend and spatial distribution of housing activeness with the help of multiple easily-accessible datasets. We first identified the regional activeness status at the individual level from energy consumption data and then matched it with nightlight and land use data geographically. Then, we introduce the principle of robustification via truncation and factor-adjusted regularization methods for prediction (FarmPredict) to deal with two important stylized features in big data. The heterogeneity of big data is mitigated through the use of the government land planning data. Farm-Predict effectively lifts the prediction space and solves the colinearity problem in high-dimensional data. It is applicable to all machine learning algorithms. FarmPredict allows us to extend the regional results to the city level, with a 75% out-of-sample explanation of the spatial and timeliness variation in the housing usage. FarmPredict is not only a model but an analytical framework of machine learning on high-dimensional data, showing broad potential applications to other social science problems. Since energy is indispensable for life, our method is highly transferable with the requirement of only public and accessible data. Our paper demonstrates the power of machine learning in understanding socioeconomic outcomes when the census and survey data is costly or unavailable.

(Joint work with Yang Zhou, Lirong Xue, Zhengyu Shi, Libo Wu)

Special Invited Speaker



Xihong Lin, Ph.D., is Professor and former Chair of the Department of Biostatistics, Coordinating Director of the Program in Quantitative Genomics at the Harvard T. H. Chan School of Public Health, and Professor of the Department of Statistics at the Faculty of Arts and Sciences of Harvard University, and Associate Member of the Broad Institute of Harvard and MIT.

Dr. Lin is an elected member of the National Academy of Medicine. She received the 2002 Mortimer Spiegelman Award from the American Public Health Association, and the 2006 Committee of Presidents of Statistical Societies (COPSS) Presidents' Award and the 2017 COPSS FN David Award. She is an elected fellow of American Statistical Association (ASA), Institute of Mathematical Statistics, and International Statistical Institute.

Dr. Lin's research interests lie in development and application of statistical and computational methods for analysis of massive data from genome, exposome and phenome, and scalable statistical inference and learning for big genomic, epidemiological and health data.

Location and Time: HPNP Auditorium (1404), June 21 (Tuesday), 13:00 pm – 13:45 pm

Organizer: Somnath Datta, Ph.D., University of Florida

Keynote Host: Ji-Hyun Lee, DrPH, University of Florida

Title: Lessons learned from the COVID-19 pandemic: a statistician's reflection

Abstract: In this article, I will discuss my experience as a statistician involved in COVID-19 research in multiple capacities in the last two years, especially in the early phase of the pandemic. I will reflect on the challenges and the lessons I have learned in pandemic research regarding data collection and access, epidemic modeling and data analysis, open science and real time dissemination of research findings, implementation science, media and public communication, and partnerships between academia, government, industry and civil society. I will also make several recommendations on preparing for the next stage of the pandemic and for future pandemics.

Special Invited Speaker



Nilanjan Chatterjee, Ph.D., received his Bachelor's and Master's degree from the Indian Statistical Institute, Calcutta and a Ph.D. in statistics from the University of Washington, Seattle in 1999. He served as Chief of the Biostatistics Branch of the Division of Cancer Epidemiology and Genetics (DCEG), National Cancer Institute (NCI) for almost eight years. Dr. Chatterjee now serves as a Bloomberg Distinguished Professor at the Johns Hopkins University with joint appointments at the Bloomberg School of Public Health (Biostatistics) and the School of Medicine (Oncology). He remains a Special Volunteer with DCEG.

Dr. Chatterjee's research focuses on a diverse set of quantitative issues that arise in design, analysis, interpretation and public health translation of modern molecular and genetic epidemiologic studies.

Location and Time: HPNP Auditorium (1404), June 21 (Tuesday), 13:45 pm – 14:30 pm

Organizer: Somnath Datta, Ph.D., University of Florida

Keynote Host: Ji-Hyun Lee, DrPH, University of Florida

Title: Predictive model building through integration of information across disparate data sources and summary-statistics

Abstract: Model building based on classical statistical methods, as well as modern machine learning techniques, typically requires availability of a single adequately large dataset, or multiple harmonized datasets across a group of similar studies. In the future, however, development of complex models incorporating a variety of factors from different domains will require integration of information from disparate data sources, which, individually may have information only on subsets of the explanatory variables of interest. Moreover, information from some studies may only be available through pre-computed summary-statistics, generated under certain forms of "reduced" models. In this talk, I will describe some of our recent efforts towards developing statistical methods for model building through data integration under a semiparametric generalized meta-analysis framework. I will illustrate the unique opportunity data integration methods provide through an application involving the development of a COVID-19 mortality risk calculator through integration of information across diverse datasets.

Banquet Speaker



Lee-Jen Wei, Ph.D., was graduated from Fu Jen Catholic University's Mathematics Department in 1970. He obtained his PhD from the University of Wisconsin–Madison in 1975. He has been a tenured Professor of Biostatistics at Harvard University since 1991 and was the co-director of the Bioinformatics Core at the Harvard School of Public Health from 2003 to 2007. From 2003 to 2004, he served as the acting chair of the Department of Biostatistics at Harvard University. Under his supervision, the department successfully converted the doctor of science degree program in biostatistics (a professional degree) to a conventional (art and sciences) Ph.D. program at the Harvard Graduate School. This was an important accomplishment since the department had tried this conversion for more than 20 years without success.

Professor Wei has developed and published a number of novel quantitative methods for analyzing data from experimental and observational studies. Specifically, he has published many papers on monitoring drug and device safety and related topics. The resulting procedures have been utilized for various drug and device regulatory evaluations involving safety issues. His extensive experience in quantitative science for making inferences about the drug and device safety is readily applicable to the general industry product safety issues.

Location and Time: Ben Hill Griffin Stadium Champions Club (121 Gale Lemerand Drive), June 21 (Tuesday), 20:00 pm – 20:45 pm

Organizer: ICSA special lecture committee

Keynote Lecture Host: Samuel Wu, Ph.D., University of Florida

Title: Lost in translation

Abstract: One of the main goals of conducting a clinical, comparative study is to obtain robust, clinically interpretable treatment effect estimates with respect to harm-benefit perspectives at the patient's level via efficient and reliable quantitative procedures. To accomplish this goal, it is important to know how to effectively translate new developments in basic data science research into clinical research and practice. Unfortunately, some commonly used statistical procedures are not translational. That is, results of the analysis may be misinterpreted or difficult to comprehend. A notorious example is use of the p-value for clinical decision making, which is not an appropriate quantifier for assessing the clinical utility of a new therapy or strategy. In this talk, we will discuss several translational problems and present possible remedies.

Keynote Speaker



Susan Murphy, Ph.D., is the Mallinckrodt Professor of Statistics and of Computer Science, Radcliffe Alumnae Professor at the Radcliffe Institute, Harvard University. Her research focuses on improving sequential, individualized, decision making in health, in particular on clinical trial design and data analysis to inform the development of mobile health treatment policies. Susan is a Fellow of the Institute of Mathematical Statistics, a Fellow of the College on Problems in Drug Dependence, a former editor of the *Annals of Statistics*, a member of the US National Academy of Medicine and a 2013 MacArthur Fellow.

Location and Time: HPNP Auditorium (1404), June 22 (Wednesday), 8:30 am – 9:30 am

Organizer: ICSA special lecture committee

Keynote Lecture Host: Guogen Shan, Ph.D., University of Florida

Title: Inference for longitudinal data after adaptive sampling

Abstract: Adaptive sampling methods, such as reinforcement learning (RL) and bandit algorithms, are increasingly used for the real-time personalization of interventions in digital applications like mobile health and education. As a result, there is a need to be able to use the resulting adaptively collected user data to address a variety of inferential questions, including questions about timevarying causal effects. However, current methods for statistical inference on such data (a) make strong assumptions regarding the environment dynamics, e.g., assume the longitudinal data follows a Markovian process, or (b) require data to be collected with one adaptive sampling algorithm per user, which excludes algorithms that learn to select actions using data collected from multiple users. These are major obstacles preventing the use of adaptive sampling algorithms more widely in practice. In this work, we proved statistical inference for the common Z-estimator based on adaptively sampled data. The inference is valid even when observations are non-stationary and highly dependent over time, and (b) allow the online adaptive sampling algorithm to learn using the data of all users. Furthermore, our inference method is robust to miss-specification of the reward models used by the adaptive sampling algorithm. This work is motivated by our work in designing the Oralalytics oral health clinical trial in which an RL adaptive sampling algorithm will be used to select treatments, yet valid statistical inference is essential for conducting primary data analyses after the trial is over.

ICSA Applied Statistics Symposium Student Paper Awards

- Margaret Banker, University of Michigan
Title: Supervised Learning of Physical Activity Features from Functional Accelerometer Data
See session 5H in the program
- Ying Jin, Stanford University
Title: Sensitivity Analysis under the f-Sensitivity Models: A Distributionally Robust Optimization Viewpoint
See session 5H in the program
- Shuting Shen, Harvard University
Title: Fast Distributed Principal Component Analysis for Large-Scale Federated Data
See session 5H in the program
- Xiulin Xie, University of Florida
Title: High-Dimensional Dynamic Process Monitoring By PCA-Based Sequential Learning
See session 5H in the program

Jiann-Ping Hsu Pharmaceutical and Regulatory Sciences Student Paper Award

- Kan Chen, University of Pennsylvania
Title: Covariate-Balancing-Aware Interpretable Deep Learning models for Treatment Effect Estimation
See session 3H in the program

SC01: Causal Inference with R

Location and Time: HPNP G114, Sun, June 19, 8:30 - 12:30

Length: Half-day

Instructors: Prof. Babette Brumback (University of Florida)

Abstract: One of the primary motivations for clinical trials and observational studies of humans is to infer cause and effect. Disentangling causation from confounding is of utmost importance. Causal Inference with R explains and relates different methods of confounding adjustment in terms of potential outcomes and graphical models, including standardization, doubly robust estimation, difference-in-differences estimation, and instrumental variables estimation. Several real data examples, simulation studies, and analyses using R motivate the methods throughout. The course assumes familiarity with basic statistics and probability, regression, and R. The course will be taught with a blend of lecture and worked examples.

Teaching Plan:

First part:

Introduction – 15 minutes

Potential Outcomes and Effect Measures – 30 minutes

Causal Directed Acyclic Graphs – 1 hr

15 minute break

Second part:

Standardization and Doubly Robust Estimation – 1 hr

Difference-in-Differences Estimation – 30 minutes

Instrumental Variables Estimation – 30 minutes

About the Instructors: Babette A. Brumback, Ph.D. is Professor in the Department of Biostatistics at the University of Florida; she won the department's Outstanding Teacher Award for 2020-2021. A Fellow of the American Statistical Association, she has researched and applied methods for causal inference since 1998, specializing in methods for time-dependent confounding, complex survey samples and clustered data.

SC02: Leveraging Real-World Data in Clinical Trial Design and Analysis

Location and Time: HPNP G103, Sun, June 19, 8:30 - 17:30

Length: Full-day

Instructor: Dr. Chenguang Wang (Regeneron Pharmaceuticals, Inc.)

Abstract: The amount of real-world data (RWD) collected from sources other than protocol-driven clinical studies is increasing ultra-rapidly. The clinical evidence that can be derived from analysis of these RWD is considered as real-world evidence (RWE) that can complement the knowledge derived from traditional well-controlled clinical trials. Leveraging RWE can potentially save time and cost of the investigational

study and improve the efficiency of regulatory decision-making. Incorporating RWD in regulatory decision-making demands much more than "mixing" RWD with investigational clinical trial data. The RWD has to undergo appropriate analysis for deriving the right RWE. Moreover, such analysis has to be integrated with the design and analysis of the investigational study for regulatory decision-making. The standard clinical trial toolbox does not offer ready solutions for incorporating RWD. In this course, the instructor(s) will cover a series of methods they have developed for leveraging real-world data in clinical trial design and analysis. Their work has been recognized by the FDA and received The FDA CDRH Excellence in Scientific Research Award and The FDA Scientific Achievement Award.

Teaching Plan: In Part I of the course, we introduce a new method for proposing performance goals—numerical target values pertaining to effectiveness or safety endpoints in single-arm medical device clinical studies—by leveraging RWE. The method applies entropy balancing to address possible patient dissimilarities between the study's target patient population and existing real-world patients, and can take into account operation differences between clinical studies and real-world clinical practice.

In Part II of the course, we introduce a method that extends the Bayesian power prior approach for a single-arm study to leverage external RWD. The method uses propensity score methodology to pre-select a subset of RWD patients that are similar to those in the current study in terms of covariates, and to stratify the selected patients together with those in the current study into more homogeneous strata. The power prior approach is then applied in each stratum to obtain stratum-specific posterior distributions, which are combined to complete the Bayesian inference for the parameters of interest.

In Part III of the course, we introduce several extensions of the PS-integrated method in Part II. These extensions include 1) a frequentist PS-integrated composite likelihood approach for incorporating RWE in single-arm clinical studies; 2) leveraging multiple RWD sources in single-arm medical device clinical studies; 3) leveraging RWD for the evaluation of diagnostic tests for low prevalence diseases; 4) augmenting both arms of a randomized controlled trial by leveraging RWD; and 5) PS-integrated approach for survival analysis.

In Part IV of the course, we describe an R package, *psrwe*, that implements a PS-integrated power prior (PSPP) method, a PS-integrated composite likelihood (PSCL) method, and a PS-integrated weighted Kaplan-Meier estimation (PSKM) method for the methods in Parts II and III. Illustrative examples are provided to demonstrate each of the approaches.

In Part V of the course, we introduce a propensity score-based Bayesian non-parametric Dirichlet process mixture model that summarizes subject-level information from randomized and RWD to draw inference on the causal treatment

effect in exploratory analysis.

About the Instructor: Dr. Chenguang Wang is a Senior Director and the Head of Statistical Innovation at Regeneron. Previously, Dr. Wang was an Associate Professor with Johns Hopkins University and an FDA Mathematical Statistician. Dr. Wang has extensive experience in clinical trial design and analysis in the regulatory setting. Dr. Wang holds B.S. and M.S. degrees in Computer Science and has abundant experience developing statistical software.

SC03: Marginal Models in Analysis of Correlated Binary Data with Time-Dependent Covariates

Location and Time: HPNP G301, Sun, June 19, 8:30 - 17:30

Length: Full-day

Instructors: Prof. Jeffrey Wilson (Arizona State University); Prof. Din Chen (Arizona State University)

Abstract: This workshop is based on the book: "Marginal Models in Analysis of Correlated Binary Data with Time Dependent Covariates" co-authored by Drs. Jeffrey R. Wilson, Elsa Vazquez-Arreola, and (Din) Ding-Geng Chen, published by Springer in 2020, which is the first book to systematically introduce marginal models to analyze correlated binary data with time-dependent covariates in clinical trials and observational studies using R and SAS. This workshop provides a thorough presentation of correlated binary data with time-dependent covariate. It gives a detailed step-by-step illustration of their implementation using R and SAS. Longitudinal data or contain correlated data due to the repeated measurements on the same subject. The changing values usually consist of time-dependent covariates and their association with the outcomes present different sources of correlation. Most methods used to analyze longitudinal data would average the effects of time-dependent covariates on outcomes over time and provide a single regression coefficient per time-dependent covariate. Such an approach prevents analysts and researchers the opportunity to following the changing impact of time-dependent covariates on the outcomes. The workshop addresses such issues through the use of partitioned regression coefficients. We further use examples of correlated data with time-dependent covariate on obesity from the Add Health study and cognitive impairment diagnosis in the National Alzheimer's Coordination Center.

Teaching Plan: Morning Session (8:30am to 12:30pm):

1. Fundamentals of estimation of regression coefficients in cross-sectional data
 - a. Review of the estimation of regression models
 - b. Generalized estimating equation (GEE) and generalized linear mixed models
 - c. Generalized Method of Moments estimates;
2. Presentation on data with time-dependent covariates and discussion on the partitioned matrix.

Afternoon Session (1:30pm to 4:30pm):

3. Present correlated data with time-dependent covariates. Illustrate longitudinal data and the analysis using linear mixed models for continuous endpoints, generalized linear mixed model and GEE for categorical endpoints.
4. Bayesian analysis in this partitioned data matrix using MCMC is applied.

About the Instructors: Dr. Jeffrey Wilson is a Professor of Statistics and Biostatistics at Arizona State University. Dr. Wilson's research experience includes grants as PI and co-PI from the NIH, NSF, USDA, Arizona Department of Health Services, and the Arizona Disease Research Commission. He is presently the Statistics Associate Editor for The Journal of Minimally Invasive Gynecology and a former Chair of the Editorial Board of the American Journal of Public Health. He has published more than 85 articles in leading journals such as Statistics in Medicine, American Journal of Public Health, Journal of Royal Statistics Society, Computational Statistics, and Australian Journal of Statistics, among others. He has consulted with pharmaceutical companies and hospitals while representing them before the FDA and other federal government healthcare agencies. He has taught specialized Biostatistics classes at Mayo Clinic. He has led similar courses for Phoenix Children's Hospital, Barrow Neurological Center, St. Joseph's Hospital, and Banner Hospital. He is the former Director of the School of Health Management and Policy He is a former Director and co-Director of the Biostatistics Core in the NIH Center for Alzheimer at Arizona State University.

Dr. (Din) Ding-Geng Chen is now the executive director and professor in biostatistics at College of Health Solutions, Arizona State University. He was the Wallace H. Kuralt distinguished professor in Biostatistics at University of North Carolina-Chapel Hill, a professor in biostatistics at the University of Rochester Medical Center, the Karl E. Peace endowed eminent scholar chair and professor in biostatistics from the Jiann-Ping Hsu College of Public Health at the Georgia Southern University. Dr. Chen is an elected fellow of the American Statistical Association (ASA), an elected member of the International Statistics Institute (ISI), and a senior expert consultant for biopharmaceuticals and government agencies with extensive expertise in clinical trial biostatistics. Dr. Chen has more than 200 referred professional publications and co-authored/co-edited 33 books on biostatistics clinical trials, biopharmaceutical statistics, interval-censored survival data analysis, meta-analysis, public health statistics, statistical causal inferences; statistical methods in big-data sciences and Monte-Carlo simulation-based statistical modeling. Dr. Chen has been invited nationally and internationally to give short courses at various scientific conferences.

SC04: Statistical methods for analyzing transmission and control of infectious diseases

Location and Time: HPNP G312, Sun, June 19, 8:30 - 12:30

Length: Half-day

Instructors: Dr. Ira Longini (University of Florida); Dr. Yang Yang (University of Florida); Dr. Matt Hitchings (University of Florida)

Abstract: Application of statistical inference methods to infectious disease data is a key tool in understanding transmissibility of pathogens and the effectiveness of interventions. In this half-day course, we will learn about different sources of data that arise from passive surveillance, active case finding and clinical studies, and methods for inferring key parameters from such data. The types of data sources to be covered include epidemic curve data, household-based observational data, and data arising from serosurveillance studies. We will also cover common computational algorithms for statistical inference and a few software packages that implement these algorithms. In addition, we will briefly introduce several advances in modeling frameworks to address challenges arising from the pandemic of COVID-19. Upon completion of this course, participants will recognize the various types of infectious disease data, common models designed to analyze these data, key parameters of epidemiological importance including intervention efficacies, and promising research directions in the field of infectious disease modeling.

Teaching Plan: The course will be divided into three sessions each of 70min, with two 15-min breaks.

First session: History of infectious disease modeling; types of infectious disease data (case numbers, serology, household data including time of symptom onset) and the underlying hierarchy of information; Overview of transmission parameters of epidemiological importance such as the basic reproductive number, final attack rate, and secondary attack rate; Different measures of vaccine efficacies and effectiveness of vaccination programs.

Second session: Detail on classic models that are fitted to epidemic curve data, final size models with fixed and random infectious periods for close contact groups (e.g., households), discrete-time chain binomial models and continuous-time survival models for sequential data of symptom onsets or laboratory confirmations among close contact groups, statistical inference from serosurveillance data, and agent-based models.

Third session: Computational methods (EM and Monte Carlo EM algorithms, traditional MCMC, Approximate Bayesian Computing, Particle Filtering, and Hamiltonian Monte Carlo). We will introduce a few R packages (e.g. surveillance, transtat, serosolver) and show some data examples; recent advances in statistical transmission models to address challenges the a rose during the pandemic of COVID-19 (e.g., presymptomatic and asymptomatic infectiousness,

under-testing, delayed reporting, etc.).

About the Instructors: Dr. Ira Longini is a professor of biostatistics in the College of Public Health and Health professions as well as Emerging Pathogens Institute at the University of Florida. He works on the mathematical modeling, stochastic processes and biostatistics applied to epidemiological infectious disease problems. He has specialized in the mathematical and statistical theory of epidemics—a process that involves constructing and analyzing mathematical models of disease transmission, disease progression and the analysis of infectious disease data based on these models. In addition, he works extensively in the design and analysis of vaccine and infectious disease prevention trials and observational studies.

Dr. Yang Yang is an associate professor of biostatistics in the College of Public Health and Health professions as well as Emerging Pathogens Institute at the University of Florida. His research focuses on statistical methods for disease transmission dynamics, efficacy evaluation, missing data and surveillance bias. He also works on ecological modeling and genetic association for clinical outcomes.

Dr. Matt Hitchings is an Assistant Professor in the Department of Biostatistics at the University of Florida. His primary focus is evaluating the effectiveness of interventions against infectious disease, through clinical trials, observational studies, and development and application of mathematical models. Recently he has been conducting observational studies of vaccine effectiveness using passive surveillance data in Brazil, and developing a framework for analysis of serological data for pathogens including SARS-CoV-2 and dengue virus.

SC05: Spatial analysis with Gaussian Markov random fields

Location and Time: HPNP G114, Sun, June 19, 13:30 - 17:30

Length: Half-day

Instructors: Dr. Debashis Mondal (Washington University)

Abstract: Gaussian Markov random fields have been applied with much success to account for discrete spatial variation in both lattice and areal unit data. Applications include astronomy, agriculture, computer vision, climate studies, epidemiology, image analysis, geology and other areas of environmental science. Lattice-based Gaussian Markov random fields are extremely adaptable to swift and uncomplicated statistical computations and provide ways to develop complex and hierarchical models through local specifications, and, for these reasons, have contributed to considerable success in the analysis of spatial data. This short course gives an introduction to spatial models based on Gaussian Markov random fields. The course covers statistical computation for spatial linear mixed models, particularly, residual maximum likelihood (REML) estimation and kriging or

prediction. The course also presents statistical computation for general spatial mixed models using Markov Chain Monte Carlo (MCMC) sampling methods. Practicum sessions will introduce various R codes with applications from environmental sciences and geographical epidemiology.

The course will end with a summary of the topics and ideas covered and a list of further resources.

Teaching Plan: Lecture 1: Introduction to spatial statistics, Gaussian Markov random fields, conditionals and intrinsic autoregressions.

Lecture 2: Spatial mixed models, REML, kriging, h-likelihood and MCMC computations.

Break

Lecture 3 and 4: Statistical calculations using R-codes. Applications from environmental sciences and geographical epidemiology.

Summary and further resources.

About the Instructors: Debashis Mondal, PhD, is an associate professor in the Department of Mathematics and Statistics at Washington University in St Louis. Mondal's research interests include spatial statistics; computational science and machine learning; and applications in environmental sciences, ecology, including microbial ecology, and geographical epidemiology. Mondal won an NSF CAREER Award in 2013 and the International Indian Statistical Association's Young Researcher Award in 2015. He is also an elected member of the International Statistical Institute. Mondal earned his doctorate in statistics at the University of Washington, Seattle.

SC06: Bayesian Computational Tools for Clinical Data

Location and Time: HPNP G312, Sun, June 19, 13:30 - 17:30

Length: Half-day

Instructor: Prof. Sujit Ghosh (North Carolina State University); Dr. Amy Shi (AstraZeneca Pharmaceutical)

Abstract: The Bayesian paradigm provides a structured and practical way of expressing complicated models through a sequence of simple conditional distributions making them useful for simple to complex data structures required to address multiple phases of clinical trials, particularly for those that involves different types of data irregularities (missing values, censored data, etc.). Over the recent years there have been tremendous efforts on developing Bayesian analytics for leveraging data from sources outside of prospectively designed study, referred to as external data such as various Real-World-Data (RWD) sources, historical clinical data, and data from multiple trials within a grand hierarchical structure. Thus, development of appropriate statistical models and related inference are warranted that are not only

based on solid theoretical guarantees but also making sure that such complex models are estimable and interpretable in practical settings for modern clinical trials. Thus, one of the main goals of the proposed short course is to present the modern analytical tools that are easily accessible to practitioners by providing a glimpse of theoretical backgrounds supplemented by many practical examples derived from real case studies. This will be accomplished by illustrating numerous real-data examples (using software demos) ranging from two-arm trials to more complex hierarchical models that involves handling data irregularities commonly faced by practitioners.

Teaching Plan: The first part of the short course will begin with a brief overview of Bayesian machine learning (BML) methods for randomized controlled trials (RCTs) using various study designs including sample size determination methods. In particular, it will showcase the use of Bayesian posterior predictive methods for properly handling missing and censored data, a feature that are not readily employed my routine ML methods. The second part of the course will involve more realistic and complex models that have recently emerged in the modern era used by pharmaceutical industries and regulatory agencies, and then showcase the use of modern BML methods through various real case studies. Throughout the tutorial practical applications and worked-out examples will be emphasized without getting into the theoretical underpinnings of the methods, but relevant literature will be provided for those wishing to learn more in-depth notions of BML tools. The concepts and methods discussed will be demonstrated using the popular software packages (R and SAS) developed by the presenters, but those are implementable by any other software capable of coding Markov Chain Monte Carlo (MCMC) methods.

The two-parts of the course will consist of the following topics:

Part I - Introduction to Bayesian Methods for Clinical Trials

1. Basics of Bayesian Methods for RCTs (20min)
2. Predictive Distributions and Sample Size Determination (20min)
3. Computational Methods using Monte Carlo Methods (35min)
4. Primer on Bayesian Software (via R, Stan and SAS) (30min)

(15min break)

Part II – BML methods with real-data examples

1. Bayesian regression models using 'brms' R package (35min)
2. GLMs and Multi-level models PROC BGLIMM (40min)
3. Penalized regression models with data irregularities (30min)
4. Q&As and additional demos on demand (15min)

About the Instructor: Professor Sujit Kumar Ghosh has

over 25 years of experience in conducting, applying, evaluating and documenting statistical analysis of biomedical and environmental data. Prof. Ghosh is actively involved in teaching, supervising and mentoring graduate students at the doctoral and master levels. He has supervised over 40 doctoral graduate students and published over 125 peer-reviewed journal articles in various areas of statistics with applications in biomedical and environmental sciences, econometrics and engineering. He has recently co-authored a book (with Dr. Reich) titled "Bayesian Statistical Methods," which is being used as a textbook at several universities. Prof. Ghosh has delivered over 180 invited lectures, seminars at national and international meetings. He has also delivered several short courses and served as short-term visiting professor at several institutions in various countries. Prof. Ghosh received the International Indian Statistical Association (IISA) Young Investigator Award in 2008; was elected a Fellow of the American Statistical Association (ASA) in 2009; was elected as the

President of the NC Chapter of ASA in 2013 and also elected as the President of the IISA in 2017.

Dr. Amy Shi is currently a Statistical Science Associate Director at AstraZeneca Pharmaceutical in the Late CVRM (Cardiovascular Renal Metabolism) group. Much of her work involves with taking part in clinical trials as a statistician and researching for innovative statistical methods. Before joining AstraZeneca, she was a Principal Research Statistician Developer in the Bayesian Modeling Group at SAS from 2010 to 2021. Her job responsibility was to enhance the Bayesian capabilities of SAS software, with a focus on generalized linear mixed models, multilevel hierarchical settings, variable selection, choice modeling, and machine learning. She developed a couple of SAS Bayesian procedures (PROC BCHOICE and PROC BGLIMM) and many functional packages. Dr. Shi has a MS in Statistics from the Michigan State University and a Ph.D. in Biostatistics from the University of North Carolina at Chapel Hill.

Scientific Program (Mon, Jun. 20 - Wed, Jun. 22)

Welcome and Opening Remarks: Mon, June 20 8:00-8:30 (EDT)

Session W : Welcome and Opening Remarks

Location: HPNP Auditorium (1404)

Organizer: Symposium Organizing Committee.

Chair: Peihua Qiu, Ph.D., University of Florida.

- 8:00-8:05 Welcome - Dr. Peihua Qiu, Chair of Department of Biostatistics
- 8:05-8:10 Welcome - Dr. Michael Perri, Dean of College of Public Health and Health Professions
- 8:10-8:20 Welcome - Dr. Zhezhen Jin, President of International Chinese Statistical Association
- 8:20-8:25 Welcome - Somnath Datta, Co-chair of organizing committee
- 8:25-8:30 Opening Remarks - Ji-Hyun Lee, Chair of local committee

Plenary Keynote Talk 1: Mon, June 20 8:30-9:30 (EDT)

Session P1 : Plenary Keynote Talk 1

Location: HPNP Auditorium (1404)

Organizer: ICSA Special Lecture Committee.

Chair: Samuel Wu, Ph.D., University of Florida.

- 8:30-9:30 Change detection, estimation, and segmentation
David O. Siegmund. Stanford University

Sessions 1A-1H: Mon, June 20 10:00-11:40 (EDT)

Session 1A : Causal Inference And Its Applications

Location: HPNP G312

Organizer: Xinping Cui, University of California, Riverside, Esra Kurum, University of California, Riverside.

Chair: Xinping Cui, University of California, Riverside.

- 10:00-10:25 A causal approach to functional mediation analysis with application to a smoking cessation intervention
Donna Coffman. Temple University
- 10:25-10:50 Estimating the Average Treatment Effect in Randomized Clinical Trials with All-or-None Compliance
Zhiwei Zhang. NIH/NCI
- 10:50-11:15 Survey Weighting Strategies In Causal Mediation Analysis
Haoyu Zhou. Temple University
- 11:15-11:40 Discussion: Causal Inference and its Applications
Esra Kurum. University of California, Riverside

Session 1B : Latent Variable Models In The Data Science Era

Location: HPNP G112

Organizer: Yuqi Gu, Columbia University, Gongjun Xu, University of Michigan.

Chair: Yuqi Gu, Columbia University.

- 10:00-10:25 Identifiable Deep Generative Models via Sparse Decoding
♦ *Gemma Moran*¹, *Dhanya Sridhar*², *Yixin Wang*³ and *David Blei*¹. ¹Columbia University ²Mila and Universite de Montreal ³University of Michigan
- 10:25-10:50 Population-Level Balance in Signed Networks
♦ *Weijing Tang and Ji Zhu.* University of Michigan
- 10:50-11:15 Likelihood estimation of sparse topic distributions in topic models and its applications to Wasserstein document distance calculations
♦ *Xin Bing, Florentina Bunea, Marten Wegkamp and Seth Strimas Mackey.* Cornell University
- 11:15-11:40 High-dimensional principle component analysis with heterogeneous missingness
♦ *Ziwei Zhu*¹, *Tengyao Wang*² and *Richard Samworth*³. ¹University of Michigan, Ann Arbor ²London School of Economics ³University of Cambridge

Session 1C : Some Advances In Statistical Machine Learning

Location: HPNP G101

Organizer: Taps Maiti, Michigan State University.

Chair: Vojtech Kejzlar, Skidmore College.

- 10:00-10:25 Structurally Sparse Bayesian Neural Networks: Spike and Slab Shrinkage Priors
Sanket Jantre, ♦ *Shrijita Bhattacharya and Tapabrata Maiti.* Michigan State University
- 10:25-10:50 An Adaptive Stochastic Approximation Algorithm for Randomized Decision GAN
Faming Liang. Purdue University
- 10:50-11:15 Volcano and valley prior with adhesive shrinkage for high dimensional data
Liangliang Zhang. case western reserve university
- 11:15-11:40 Information-preserving Bayesian models for efficient and robust learning
Sandeep Madireddy. Argonne NationalLaboratory

Session 1D : Machine Learning/Artificial Intelligence In Biomedical Research With 'big' Data

Location: HPNP G103

Organizer: Xiang-Yang Lou, University of Florida/Department of Biostatistics, Qing Lu, University of Florida/Department of Biostatistics.

Chair: Xiang-Yang Lou, University of Florida/Department of Biostatistics.

- 10:00-10:25 AI for Regulatory Science
Weida Tong. FDA

10:25-10:50 Causal networks for drug discovery
♦Tao Xu¹, Shicheng Guo², Jinyung Zhao¹ and Momiao Xiong³. ¹University of Florida ²University of Wisconsin-Madison ³University of Texas Health Science Center at Houston

10:50-11:15 New Toolkits for Disease Network Biology
Jake Chen. UAB Informatics Institute

11:15-11:40 Achieving Differential Privacy with Matrix Masking in Big Data
Aidong Ding¹, ♦Samuel Wu², Guanhong Miao² and Shigang Chen². ¹Northeastern University ²University of Florida

Session 1E : Statistical Challenges And Advances In Complex Data Analysis

Location: HPNP G114

Organizer: Yichuan Zhao, Georgia State University.

Chair: Yichuan Zhao, Georgia State University.

10:00-10:25 Nontraditional Statistical Methods based on Wasserstein Distances and Conformal Prediction Set
Xiaoming Huo. Georgia Institute of Technology

10:25-10:50 Bayesian Spatially Varying Weight Neural Networks with the Soft-Thresholded Gaussian Process Prior
Jian Kang. University of Michigan

10:50-11:15 Some Recent Advances on the analysis of Interval-Censored Case-cohort Failure Time Data
(Tony) Jianguo Sun. University of Missouri

11:15-11:40 An Efficient Method for Clustering Multivariate Longitudinal Data
Junyi Zhou¹, ♦Ying Zhang² and Wanzhu Tu³. ¹Amgen Inc ²UNMC ³Indiana University

Session 1F : Statistical Methods And Applications For Analyzing Real-World Data

Location: HPNP G301

Organizer: Kelly Zou, Viatrix.

Chair: Ying Lu, Stanford University.

10:00-10:25 WeightP2V: a flexible risk prediction framework with patient representation weighted by medical concepts
Jia Guo and ♦Shuang Wang. Columbia University

10:25-10:50 Efficient Algorithms and Implementation of a Semiparametric Joint Model for Longitudinal and Competing Risks Data: With Applications to Massive Biobank Data
Shanpeng Li¹, Ning Li¹, Hong Wang², Jin Zhou¹, Hua Zhou¹ and ♦Gang Li¹. ¹UCLA ²Central South University

10:50-11:15 A statistical quality assessment method for longitudinal observations in electronic health record data with an application to the VA million veteran program
Hui Wang¹, Ilana Belitskaya-Levy¹, Fan Wu¹, Jennifer Lee², Mei-Chiung Shih¹, Philip Tsao² and ♦Ying Lu. ¹Department of Veterans Affairs, Palo Alto, CA, USA ²Stanford University

11:15-11:40 Floor Discussion.

Session 1G : Recent Advances In Survival And Recurrent Events Analysis For Complex Data Structures

Location: HPNP 1101

Organizer: Dongdong Li, Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute.

Chair: Dongdong Li, Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute.

10:00-10:25 Structured variable selection in Cox model with time-dependent covariates

♦Guanbo Wang¹, Yi Yang¹, Mirelle Schnitzer², Tom Chen³, Rui Wang³ and Robert Platt¹. ¹McGill University ²University of Montreal ³Harvard University

10:25-10:50 Robust Estimation for Recurrent Event Analysis in the Presence of Informative Event Censoring

♦Tom Chen¹, Rui Wang¹ and Victor Degruttola². ¹Harvard Pilgrim Health Care and Harvard Medical School ²Harvard School of Public Health

10:50-11:15 Variance Estimation for Cox Model When Using Propensity Score Weighting

♦Di Shu¹, Jessica G Young², Sengwee Toh² and Rui Wang². ¹University of Pennsylvania ²Harvard University

11:15-11:40 Statistical Analysis of Recurrent Events from Administrative Databases

Yi Xiong. Fred Hutchinson Cancer Center

Session 1H : Statistical Inference For Two-Phase Studies With Outcome-Dependent Sampling

Location: HPNP 1102

Organizer: Natalie DelRocco, University of Florida Department of Biostatistics.

Chair: Adam Ding, Northeastern University Department of Mathematics.

10:00-10:25 Robust methods for Two-Phase Studies under generalized linear models

♦Jacob Maronge¹, Jonathan Schildcrout² and Paul Rathouz³. ¹University of Texas MD Anderson Cancer Center ²Vanderbilt University Medical Center ³Dell Medical School at the University of Texas at Austin

10:25-10:50 Epidemiological Study Designs for Quantitative Longitudinal Data

♦Jonathan Schildcrout, Chiara Digraio and Ran Tao. VUMC

10:50-11:15 Statistical Methods for Selective Biomarker Testing in Two-Phase Studies

♦Natalie Delrocco¹, Adam Ding² and Samuel Wu¹. ¹University of Florida ²Northeastern University

11:15-11:40 Design and Analysis Strategies with "Secondary" Use Data

Sarah Lotspeich. UNC

Sessions 2B-2H: Mon, June 20 13:00-14:40 (EDT)**Session 2B : Advanced Research In Bio-Molecular And Imaging Data By Our Young Researchers**

Location: HPNP G112

Organizer: Susmita Datta, Department of Biostatistics, University of Florida.

Chair: Zhigang Li, Department of Biostatistics, University of Florida.

13:00-13:25 Outcome-guided Bayesian Clustering for Disease Subtype Discovery Using High-dimensional Transcriptomic Data
Lingsong Meng and ♦Zhiguang Huo. Department of Biostatistics, University of Florida

13:25-13:50 Double soft-thresholded multigroup model for vector-valued image regression with application to DTI imaging
♦*Arkaprava Roy¹ and Zhou Lan².* ¹University of Florida ²Yale University

13:50-14:15 Joint analysis and visualization of DNA methylation and nucleosome occupancy in single-molecule and single-cell data
Rhonda Bacher. University of Florida

14:15-14:40 Unity in diversity: Commonalities in these three different data analytical techniques
Arkaprava Roy, Rhonda L Bacher, Zhiguang Huo and ♦Susmita Datta. University of Florida

Session 2C : Emerging Topics In Statistical Learning For Biomedical Data

Location: HPNP G101

Organizer: Li-Xuan Qin, Memorial Sloan Kettering Cancer Center.

Chair: Carrie Wright, Johns Hopkins University.

13:00-13:25 A Semiparametric Approach to Developing Well-calibrated Models for Predicting Binary Outcomes
♦*Yaqi Cao¹, Ying Yang² and Jinbo Chen¹.* ¹University of Pennsylvania ²Tsinghua University

13:25-13:50 How does data preprocessing impact statistical learning in microRNA studies?
Li-Xuan Qin. MSKCC

13:50-14:15 A Bayesian Reinforcement Learning Approach for Optimizing Combination Antiretroviral Therapy in People with HIV
♦*Yanxun Xu¹, Wei Jin¹, Yang Ni² and Leah Rubin¹.* ¹Johns Hopkins University ²Texas A&M University

14:15-14:40 HID machine: A Random Forest-based High Order Interaction Discovery Method for High-Dimensional Genomic Data
♦*Min Lu, Yifan Sha and Xi Chen.* University of Miami

Session 2D : Statistics In Biosciences (Sibs): Real World Challenges And Recent Methodological Developments

Location: HPNP G103

Organizer: X. Joan Hu, Department of Statistics and Actuarial Science Simon Fraser University, Hongzhe Lee, University of Pennsylvania.

Chair: Hongkai Ji, Johns Hopkins Bloomberg School of Public Health.

13:00-13:25 Multi-sample single-cell RNA-seq data analysis and visualization - methods, software, and benchmark

♦*Hongkai Ji¹, Boyang Zhang¹, Wenpin Hou¹, Zhicheng Ji², Zeyu Chen³, E John Wherry³ and Stephanie Hicks.* ¹Johns Hopkins Bloomberg School of Public Health ²Duke University School of Medicine ³University of Pennsylvania Perelman School of Medicine

13:25-13:50 An efficient segmentation algorithm to estimate sleep duration from actigraphy data

Jonggyu Baek¹, ♦Margaret Banker², Erica Jensen², Xichen She², Karen Peterson², Andrew Pitchford³ and ♦Peter Song. ¹University of Massachusetts Medical School ²University of Michigan ³Iowa State University

13:50-14:15 Semiparametric estimation for length-biased interval-censored data with a cure fraction

Pao-Sheng Shen¹, ♦Yingwei Peng², Hsin-Jen Chen³ and Chyong-Mei Chen³. ¹Tunghai University ²Queen's University ³National Yang Ming Chiao Tung University

14:15-14:40 Floor Discussion.

Session 2E : Some Recent Methods For Sequential Monitoring Of Complex Data

Location: HPNP G114

Organizer: Peihua Qiu, University of Florida.

Chair: Peihua Qiu, University of Florida.

13:00-13:25 A Robust Dynamic Screening System By Estimation of the Longitudinal Data Distribution

♦*Lu You¹ and Peihua Qiu².* ¹University of South Florida ²University of Florida

13:25-13:50 Transparent Sequential Learning for Statistical Process Control

Peihua Qiu. Founding Chair, Department of Biostatistics

13:50-14:15 Statistical Quality Control Using Image Intelligence: A Sparse Learning Approach

Yicheng Kang. Bentley University

14:15-14:40 Adaptive Process Monitoring Using Covariate Information

♦*Kai Yang¹ and Peihua Qiu².* ¹Medical College of Wisconsin ²University of Florida

Session 2F : Big Data, Machine Learning And Graphical Methods

Location: HPNP G301

Organizer: Kelly Zou, Viatrix.

Chair: Yuqi Gu, Columbia University.

13:00-13:25 A Latent State Space Model for Learning Brain Dynamics for Mental Disorders

Yuanjia Wang. Columbia University

13:25-13:50 Clinical practice management of primary open-angle glaucoma in the United States: An analysis of real-world evidence

Joseph Imperato¹, Kelly Zou², Jim Li² and ♦Tarek Hassan³. ¹IQVIA ²Medical Analytics and Real-World Evidence, Viatrix Inc ³Global Therapeutic Area Lead, Ophthalmology, Viatrix Inc

13:50-14:15 Bayesian Pyramids: Identifiable Multilayer Discrete Latent Structure Models for Discrete Data
♦Yuqi Gu¹ and David Dunson². ¹Columbia University
²Duke University

14:15-14:40 Role of AI/ML and Big Data Analytics in Drug and Digital Medicine Development
Peter Zhang. Otsuka Pharmaceuticals (US)

Session 2G : Recent Development In Survival Analysis In Clinical Trials

Location: HPNP 1101

Organizer: Tianmeng Lyu, Novartis, Dong Xi, Gilead Sciences.

Chair: Dong Xi, Gilead Sciences.

13:00-13:25 On the Use of Restricted Mean Survival Time in Time-to-Event Data Analysis
Lihui Zhao. Northwestern University

13:25-13:50 From Logic-respecting Efficacy Estimands to Logic-ensuring Analysis Principle for Time-to-event Endpoint in Randomized Clinical Trials with Subgroups
Yi Liu¹, Miao Yang¹, Siyoen Kil², Jiang Li³, Shoubhik Mondal⁴, Hong Tian³, Liwei Wang, ♦Yue Shentu⁵ and Godwin Yung⁶. ¹Nektar Therapeutics ²LSK ³Beigene ⁴AstraZeneca ⁵Daiichi Sankyo Inc. ⁶Genentech

13:50-14:15 A MCP-Mod approach to designing and analyzing survival trials with potential non-proportional hazards
♦Xiaodong Luo, Yuan Sun and Zhixing Xu. Sanofi

14:15-14:40 Bayesian inference for a principal stratum estimand on recurrent events truncated by death
♦Tianmeng Lyu, Björn Bornkamp, Guenther Mueller-Velten and Heinz Schmidli. Novartis

Session 2H : Challenges And Recent Developments In Multi-Outcome Analysis

Location: HPNP 1102

Organizer: Ming Wang, Penn State College of Medicine.

Chair: Ming Wang, Penn State College of Medicine.

13:00-13:25 Alternative multivariate endpoints and related statistical models for clinical trials in Alzheimer disease
Guoqiao Wang. Division of Biostatistics, Washington University in St Louis

13:25-13:50 Joint multivariate copula-frailty modeling of multiple-type recurrent events and the terminal event
Menglu Liang and ♦Ming Wang. Penn State College of Medicine

13:50-14:15 Knowledge-guided Bayesian Factor Analysis for Integrative Analysis of Multi-Omics Data
Qiyiwen Zhang, Changgee Chang and ♦Qi Long. University of Pennsylvania

14:15-14:40 Synergistic Self-learning Approach to Establishing Individualized Treatment Rules from Multiple Benefit Outcomes in a Calcium Supplementation Trial
♦Yiwang Zhou¹ and Peter Song². ¹Department of Biostatistics, St. Jude Children's Research Hospital ²Department of Biostatistics, University of Michigan

Sessions 3A-3H: Mon, June 20, 15:00-16:40 (EDT)

Session 3A : Recent Advances In Statistical Methods For Causal Inference And Personalized Medicine

Location: HPNP G312

Organizer: Ming Wang, Penn State College of Medicine.

Chair: Ming Wang, Penn State College of Medicine.

15:00-15:25 Evaluating different methods for estimating optimal treatment based on observational data

Qian Xu, Qi Zheng and ♦Maiying Kong. University of Louisville

15:25-15:50 Evidence factors from multiple, possibly invalid, instrumental variables

Anqi Zhao¹, ♦Youjin Lee², Dylan Small³ and Bikram Karmakar⁴. ¹National University of Singapore ²Brown University ³University of Pennsylvania ⁴University of Florida

15:50-16:15 Estimation of marginal treatment effect on binary outcome with multiple robustness and information borrow from secondary outcomes

♦Chixiang Chen¹, Shuo Chen¹, Qi Long², Sudeshna Das³ and Ming Wang⁴. ¹University of Maryland, School of Medicine ²University of Pennsylvania ³Harvard Medical School ⁴Pennsylvania State University

16:15-16:40 Estimation Of Marginal Treatment Effect On Binary Outcome With Multiple Robustness And Information Borrow From Secondary Outcomes

Xiasuan Cai¹, Xinru Wang¹, Justin Baker², Jukka-Pekka Onnela³ and ♦Linda Valeri¹. ¹Columbia University ²McLean Hospital ³Harvard University

Session 3B : New Advances In High-Dimensional Data Analysis

Location: HPNP G112

Organizer: Arkaprava Roy, University of Florida.

Chair: Arkaprava Roy, University of Florida.

15:00-15:25 On Statistical Inference with High Dimensional Sparse CCA
♦Nilanjana Laha¹, Nathan Huey¹, Brent Coull² and Ra-harshi Mukherjee¹. ¹Harvard University ²Harvard Mukherjee

15:25-15:50 A Graphical Lasso model for Hermitian matrices to detect global time-lagged Teleconnections

♦Indranil Sahoo¹, Joseph Guinness² and Brian J. Reich³. ¹Virginia Commonwealth University ²Cornell University ³North Carolina State University

15:50-16:15 Multilayer Adjusted Cluster Point Process Model: Application to Microbial Biofilm Image Data Analysis

♦Suman Majumder¹, Brent Coull¹, Jessica Markwelch², Floyd Dewhirst³, Jacqueline Starr⁴ and Kyu Ha Lee¹. ¹Harvard T.H. Chan School of Public Health ²Marine Biological Laboratories ³Forsyth Institute ⁴Brigham and Women's Hospital

16:15-16:40 Correlated Wishart Matrices Classification via an Expectation-Maximization Composite Likelihood-Based Algorithm
Zhou Lan. Yale University

Organizer: Zhiguang Huo, Department of Biostatistics, University of Florida.
 Chair: Zhiguang Huo, Department of Biostatistics, University of Florida.

Session 3C : Machine Learning And Deep Learning Methods For Complex And Big Data

Location: HPNP G101

Organizer: Yichuan Zhao, Georgia State University.

Chair: Yichuan Zhao, Georgia State University.

15:00-15:25 Generative models for diabetic retinopathy
Lingsong Zhang. Purdue University

15:25-15:50 Divide and conquer approaches for nonparametric regression and variable selection
Sapuni Chandrasena and ♦Rong Liu. University of Toledo

15:50-16:15 A Bayesian Semi-supervised Approach to Keyword Extraction with Only Positive and Unlabeled Data
Guanshen Wang¹, ♦Yichen Cheng², Yusen Xia², Qiang Lin³ and Xinlei Wang¹. ¹Southern Methodist University ²Georgia State University ³University of Science and Technology of China

16:15-16:40 Deep learning approaches for predicting virus-host interactions and drug response
Zhongming Zhao. University of Texas Health Science Center at Houston

Session 3D : Advance In Statistical Methods For Complex Data

Location: HPNP G103

Organizer: Dehan Kong, University of Toronto.

Chair: Dehan Kong, University of Toronto.

15:00-15:25 Predicting long-term breast cancer risk with mammogram imaging data
 ♦*Shu Jiang¹, Jiguo Cao², Bernard Rosner³ and Graham Colditz¹.* ¹Washington university school of medicine ²Simon fraser university ³Harvard School of medicine

15:25-15:50 Fighting Noise with Noise: Causal Inference with Many Candidate Instruments
 ♦*Xinyi Zhang, Linbo Wang, Stanislav Volgushev and Dehan Kong.* University of Toronto

15:50-16:15 Smooth nonparametric dynamic prediction for competing risks via deep learning
Zhiyang Zhou. University of Manitoba

16:15-16:40 Distributed Cox Proportional Hazards Model Using Summary-level Information
 ♦*Dongdong Li¹, Wenbin Lu², Di Shu³, Sengwee Toh¹ and Rui Wang⁴.* ¹Harvard Medical School ²North Carolina State University ³University of Pennsylvania Perelman School of Medicine ⁴Harvard Medical School and Harvard T.H. Chan School of Public Health

Session 3E : Recent Advancement In Statistical Learning Methods For High-Dimensional Biomedical Data

Location: HPNP G114

15:00-15:25 On p-value combination of independent and frequent signals: asymptotic efficiency and Fisher ensemble
 ♦*Yusi Fang¹, Chung Chang² and George Tseng¹.* ¹Biostatistics, University of Pittsburgh ²Applied Math, National Sun Yat-sen University

15:25-15:50 Improve Health Equality for Polygenic Risk Score (PRS) by Joint Penalized Regression of GWAS Summary Statistics from Two Ancestries
 ♦*Peng Liu¹, Max G'sell¹, Bernie Delvin² and Kathryn Roeder¹.* ¹Carnegie Mellon University ²University of Pittsburgh

15:50-16:15 High-dimension to high-dimension screening for detecting genome-wide epigenetic regulators of gene expression
Hongjie Ke¹, Zhao Ren², Shuo Chen¹, George Tseng², Jianfei Qi¹ and ♦Tianzhou Ma¹. ¹University of Maryland ²University of Pittsburgh

16:15-16:40 The mediating role of neuroimaging data in age-related cognitive decline
 ♦*Hwiyoung Lee and Shuo Chen.* University of Maryland, Baltimore

Session 3F : Advanced Statistical Learning Methods For Dynamic Systems

Location: HPNP G301

Organizer: Rongjie Liu, Florida State University.

Chair: Rongjie Liu, Florida State University.

15:00-15:25 A Computing Algorithm for Parameter Estimation of Ultra-high Dimensional VAR Model
Hongyu Miao. Florida State University

15:25-15:50 Generalized Ordinary Differential Equation (GODE) Model and Its Link to Deep Learning
Hulin Wu. University of Texas Health Science Center at Houston

15:50-16:15 Nonparametric Bayesian Q-learning for adjusting partial compliance in multi-stage randomized trials
 ♦*Indrabati Bhattacharya, Brent Johnson and Ashkan Ertefaie.* University of Rochester

16:15-16:40 Dynamic Topological Data Analysis for Brain Networks
Moo Chung¹. University of Wisconsin-Madison

Session 3G : Geometric Statistics In Medical Image Computing

Location: HPNP 1101

Organizer: Hani Doss, University of Florida.

Chair: Hani Doss, University of Florida.

15:00-15:25 Statistical Analysis of Shape Networks
 ♦*Anuj Srivastava, Xiaoyang Guo, Aditi Basu Bal and Tom Needham.* Florida State University

15:25-15:50 Feature Gradient Flow for Interpretation of Deep Learning Models
P. Thomas Fletcher. University of Virginia

- 15:50-16:15 Nested Homogeneous Spaces: Construction, Learning and Applications
Baba Vemuri. University of Florida
- 16:15-16:40 Integrated Construction of Multimodal Atlases with Structural Connectomes in the Space of Riemannian Metrics
Sarang Joshi. University of Utah

Session 3H: The Jiann-Ping Hsu Invited Session on Biostatistical and Regulatory Sciences

Location: HPNP 1102

Organizer: Lili Yu and Karl Peace, JPH College of Public Health, Georgia Southern University.

Chair: Lili Yu and Karl Peace, JPH College of Public Health, Georgia Southern University.

- 15:00-15:25 Covariate-Balancing-Aware Interpretable Deep Learning Models for Treatment Effect Estimation
♦*Kan Chen, Qishuo Yin and Qi Long*. University of Pennsylvania
- 15:25-15:50 How to Implement the “One Patient, One Vote” Principle under the Framework of Estimand?
♦*Naitee Ting*. Boehringer Ingelheim
- 15:50-16:15 Cox Model for Weibull Survival Data
Mario Keko, ♦Marwan Alsharman, Djhenne Dalmacy, Lili Yu. Georgia Southern University
- 16:15-16:40 An Application of the Cure Model to A Cardiovascular Clinical Trial
♦*Varadan Sevilimedu, S Ma, P Hartigan, TC Kyriakides*. Memorial Sloan Kettering Cancer Center

Sessions 4A-4H: Mon, June 20, 17:00-18:40 (EDT)

Session 4A : Recent Developments For Causal Inference: Theory, Method, And Application (This session is co-sponsored by the Caucus for Women in Statistics (CWS))

Location: HPNP G312

Organizer: Guanyu Hu, University of Missouri.

Chair: Guanyu Hu, University of Missouri.

- 17:00-17:25 Calibrated Optimal Decision Making with Multiple Data Sources and Limited Outcome
♦*Hengrui Cai, Wenbin Lu and Rui Song*. North Carolina State University
- 17:25-17:50 A Focusing Framework for Testing Bi-Directional Causal Effects with GWAS Summary Data
Ting Ye. University of Washington
- 17:50-18:15 Sensitivity Analysis of Individual Treatment Effects: A Robust Conformal Inference Approach
*Ying Jin*¹, ♦*Zhimei Ren*² and *Emmanuel Candès*¹.
¹Stanford University ²University of Chicago
- 18:15-18:40 Causal inference of time-varying effects in non-stationary time series using mobile health data
♦*Xiaoxuan Cai*¹, *Jukka-Pekka Onnela*², *Justin Baker*³, *Habib Rahimi-Eichi*³ and *Linda Valeri*¹. ¹Columbia University ²Harvard University ³McLean Hospital

Session 4B : High-Dimensional Statistical Inference For Big Complicated Data

Location: HPNP G112

Organizer: Gongjun Xu, Department of Statistics, University of Michigan, Yinqiu He, Data Science Institute, Columbia University.

Chair: Yinqiu He, Data Science Institute, Columbia University.

- 17:00-17:25 Anti-Concentration of Suprema of Gaussian Processes with Applications to High-Dimensional CLTs
Alexander Giessing. University of Washington
- 17:25-17:50 Multiple-Splitting Projection Test for High-Dimensional Mean Vectors
*Wanjun Liu*¹, ♦*Xiufan Yu*² and *Runze Li*³. ¹LinkedIn Corporation ²University of Notre Dame ³Penn State University
- 17:50-18:15 Two-sample hypothesis testing of multiple-network data
♦*Yinqiu He*¹, *Xuming He*², *Ji Zhu*² and *Gongjun Xu*².
¹Columbia University ²University of Michigan
- 18:15-18:40 Doubly Debiased Lasso: High-Dimensional Inference under Hidden Confounding
♦*Zijian Guo*¹, *Domagoj Cevic*² and *Peter Buhlmann*².
¹Rutgers ²ETH, Zurich

Session 4C : New Fronts In Joint Modeling And Machine Learning

Location: HPNP G101

Organizer: Zhigang Li, University of Florida.

Chair: Lihui Zhao, Northwestern University.

- 17:00-17:25 Joint modeling for longitudinal and interval censored survival data
Ding-Geng Chen. Arizona State University
- 17:25-17:50 Heterogeneous Data Integration And The Predictive Ability of Cancer Survival Models
Yi Guo. Health Outcomes & Biomedical Informatics, University of Florida
- 17:50-18:15 Regression Analysis of Mixed Panel-Count Data with Application to Cancer Studies
♦*Yimei Li*¹, *Liang Zhu*², *Lei Liu*³ and *Leslie Robison*⁴. ¹St Jude Children’s Research Hospital ²Eisai ³Washington University ⁴St. Jude Children’s Research Hospital
- 18:15-18:40 Joint modeling in presence of informative censoring in palliative care studies
♦*Quran Wu*¹, *Michael Daniels*², *Areej Jawahri*³, *Marie Bakitas*⁴ and *Zhigang Li*¹. ¹Department of Biostatistics, University of Florida ²Department of Statistics, University of Florida ³Department of Oncology, Massachusetts General Hospital ⁴School of Nursing, University of Alabama at Birmingham

Session 4D : Knowledge-Guided Machine Learning And Statistical Modeling In Longitudinal Studies With Survival Endpoints

Location: HPNP G103

Organizer: Colin Wu, National Heart, Lung and Blood Institute, Xin Tian, National Heart, Lung and Blood Institute.

Chair: Xin Tian, National Heart, Lung and Blood Institute.

- 17:00-17:25 Design and Analysis of a Multi-Platform Trial of Patients Hospitalized for COVID-19
♦ *Eric Leifer*¹, *Lucy Kornblith*, *Jeffrey Berger*, *Lana Castellucci*, *Michael Farkouh*, *Ewan Goligher*, *Patrick Lawler* and *Scott Berry*. ¹NIH/NHLI
- 17:25-17:50 Knowledge-Guided Model Building and Estimation with Time-to-Event Outcomes and Longitudinal Covariates
♦ *Colin O. Wu*¹, *Xiaoyang Ma* and *Xin Tian*. ¹Division of Intramural Research
- 17:50-18:15 Dynamic Risk Prediction Triggered by Intermediate Events Using Survival Tree Ensembles
♦ *Yifei Sun*¹, *Sy Han Chiou*², *Colin Wu*³, *Meghan McGarry*⁴ and *Chiung-Yu Huang*⁴. ¹Columbia University ²University of Texas at Dallas ³National Heart, Lung, and Blood Institute ⁴University of California San Francisco
- 18:15-18:40 Dealing With Competing Risks in Clinical Trials
James Troendle. NIH
- Session 4E : Robust Information Integration From Multiple Studies In Clinical And Biomedical Research**
Location: HPNP G114
Organizer: Ming Wang, Pennsylvania State University, Chixiang Chen, University of Maryland.
Chair: Chixiang Chen, University of Maryland.
- 17:00-17:25 On multi-site collaboration, data sharing, and analytic strategy in medical research
♦ *Jing Huang*¹, *Rui Duan*² and *Yong Chen*¹. ¹University of Pennsylvania ²Harvard University
- 17:25-17:50 Integrating summary information from many external studies with heterogeneous populations
Peisong Han. University of Michigan
- 17:50-18:15 Data Integration Methods Targeting Underrepresented Populations in Precision Medicine
Rui Duan. Harvard University
- 18:15-18:40 Integrated Analysis of Randomized Clinical Trials with Real-World Data
♦ *Xiaofei Wang*¹, *Dasom Lee*² and *Shu Yang*². ¹Duke University ²NC State University
- Session 4F : Statistical Innovation In Complex And High Dimensional Data**
Location: HPNP G301
Organizer: Jiaying Weng, Bentley University.
Chair: Zi Ye, Lehigh University.
- 17:00-17:25 Change detection in certain random intensity-driven point processes through repeated testing
Moinak Bhaduri. Bentley University
- 17:25-17:50 A nonparametric multi-sample test for high-dimensional compositional data with applications to the human microbiome
Qingyang Zhang. University of Arkansas
- 17:50-18:15 Minimum discrepancy approach for dimension reduction by filtered feature
Pei Wang. Miami University
- 18:15-18:40 Nonparametric Mixture Model: Application in Contaminated Trials
Zi Ye. Lehigh University
- Session 4G : Enhance Decision Making In Early Oncology Studies To Expedite Drug Development**
Location: HPNP 1101
Organizer: Gaohong Dong, BeiGene.
Chair: Kathy Zhang, BeiGene.
- 17:00-17:25 A Bayesian hierarchical monitoring design for phase II cancer clinical trials: Incorporating information on response duration
♦ *Jian Wang*¹, *Jing Ning*¹, *Junsheng Ma*¹, *Chunyan Cai*² and *Naval Daver*¹. ¹The University of Texas MD Anderson Cancer Center ²Marketplace Data Science, Uber
- 17:25-17:50 Bayesian Interim Monitoring for Faster Decision-Making in Early Phase Trials
Victoria Chang, *Kathy Zhang* and ♦ *Gaohong Dong*. BeiGene
- 17:50-18:15 Discussant: Ying Lu.
- 18:15-18:40 Floor Discussion.
- Session 4H : Design And Analysis Of Computer Experiments**
Location: HPNP 1102
Organizer: Abhyuday Mandal, University of Georgia.
Chair: Ting Zhang, University of Georgia.
- 17:00-17:25 Modeling and Active Learning for Experiments with Quantitative-Sequence Factors
Abhyuday Mandal. University of Georgia
- 17:25-17:50 Lioness Algorithm for Finding Optimal Design of Experiments
♦ *Hongzhi Wang*, *Qian Xiao* and *Abhyuday Mandal*. University of Georgia
- 17:50-18:15 A Simulation Optimization Approach for Sequential Accelerated Life Testing via Approximate Bayesian Inference
*Ye Chen*¹, ♦ *Qiong Zhang*², *Mingyang Li*³ and *Wenjun Cai*⁴. ¹Virginia Commonwealth University ²Clemson University ³USF ⁴Virginia Tech
- 18:15-18:40 Optimal Crossover Designs for Generalized Linear Models
♦ *Jeevan Jankar*¹, *Abhyuday Mandal* and *Jie Yang*². ¹University of Georgia ²University of Georgia
- Plenary Keynote Talk 2: Tue, June 21, 8:30-9:30 (EDT)**
- Session P2 : Plenary Keynote Talk 2**
Location: HPNP Auditorium (1404)
Organizer: ICSA Special Lecture Committee.
Chair: Somnath Datta, Ph.D., University of Florida.
- 8:30-9:30 Measuring housing activeness from multi-source big data and machine learning
Jianqing Fan. Princeton University

Sessions 5A-5H: Tue, June 21, 10:00-11:40 (EDT)**Session 5A : Statistical Methodologies In Causal Inference With Application In Drug Development**

Location: HPNP G312

Organizer: Jiarui Lu, Novartis Pharmaceuticals Corporation, Dong Xi, Gilead Sciences.

Chair: Tianmeng Lyu, Novartis Pharmaceuticals Corporation.

10:00-10:25 Time and Causality: Learning Causal Structures from Longitudinal Data

♦*Siyi Deng*¹, ♦*Jiarui Lu*² and ♦*Dong Xi*³. ¹Cornell University ²Novartis pharmaceuticals corporation ³Gilead Sciences

10:25-10:50 Minimax optimal subgroup identification

♦*Matteo Bonvini*¹, ♦*Edward H. Kennedy*¹ and ♦*Luke J. Keele*². ¹Carnegie Mellon University ²University of Pennsylvania

10:50-11:15 A Bayesian Machine Learning Approach for Estimating Heterogeneous Survivor Causal Effects: Applications to a Critical Care Trial

♦*Xinyuan Chen*¹, ♦*Michael O. Harhay*², ♦*Guangyu Tong*³ and ♦*Fan Li*³. ¹Mississippi State University ²University of Pennsylvania ³Yale University

11:15-11:40 Application of the causal inference in estimands for a principal stratum in clinical trials

♦*Yongming Qu*. Eli Lilly and Company**Session 5B : Recent Developments Of Dimension Reduction In Integrating Big And Complex Data**

Location: HPNP G112

Organizer: Zhihua Su, University of Florida.

Chair: Zhihua Su, University of Florida.

10:00-10:25 Nonlinear envelope model

♦*Bing Li*¹, ♦*Zhihua Su*² and ♦*Dennis Cook*³. ¹Penn State University ²University of Florida ³University of Minnesota

10:25-10:50 Asymptotic distribution for partial least square prediction when the number of sample is small

♦*Liliana Forzani*¹ and ♦*R. Dennis Cook*². ¹Universidad Nacional del Litoral ²University of Minnesota

10:50-11:15 A unified framework to high dimensional sufficient dimension reduction

♦*Shanshan Ding*¹, ♦*Wei Qian*¹ and ♦*Lan Wang*². ¹University of Delaware ²University of Miami

11:15-11:40 Envelope-based Partial Least Squares with Application to Cytokine-based Biomarker Analysis for COVID-19

♦*Yeonhee Park*¹, ♦*Zhihua Su*² and ♦*Dongjun Chung*³. ¹University of Wisconsin ²University of Florida ³Ohio State University**Session 5C : Precision Digital Health Care Via Machine Learning (This session is co-sponsored by the Statistical Learning and Data Science (SLDS) Section of ASA)**

Location: HPNP G101

Organizer: Glen Wright Colopy, LifeBell AI / ASA SL&DS Section Program Chair.

Chair: Samaneh Nasiri, Harvard Medical School.

10:00-10:25 Designing Reinforcement Learning Algorithms for Digital Interventions: Pre-implementation Guidelines

♦*Anna L. Trella*¹, ♦*Kelly W. Zhang*¹, ♦*Inbal Nahum-Shani*², ♦*Vivek Shetty*³, ♦*Finale Doshi-Velez*¹ and ♦*Susan A. Murphy*¹.¹Harvard University ²University of Michigan ³University of California, Los Angeles

10:25-10:50 Oblique random survival forests version 2.0: faster and more interpretable

♦*Byron Jaeger* and ♦*Nicholas Pajewski*. Wake Forest School of Medicine

10:50-11:15 Going Beyond Spike-and-slab: L1-ball Sparsity Prior With Applications On Image Data Analysis

♦*Leo Duan* and ♦*Maoran Xu*. University of Florida

11:15-11:40 Floor Discussion.

Session 5D : Statistical Methods For Complex And High Dimensional Data

Location: HPNP G103

Organizer: Xueying Tang, University of Arizona.

Chair: Xueying Tang, University of Arizona.

10:00-10:25 Consistent and scalable Bayesian joint variable and graph selection for disease diagnosis leveraging functional brain network

♦*Xuan Cao*¹ and ♦*Kyoungjae Lee*². ¹University of Cincinnati ²Sungkyunkwan University

10:25-10:50 Bayesian mixture models, non-local prior formulations and MCMC algorithms

♦*Jairo Alberto Fuquenepatino*. UC Davis

10:50-11:15 Two-component Gibbs samplers: Convergence rate and asymptotic variance

♦*Qian Qin*¹ and ♦*Galin Jones*². ¹University of Minnesota ²University of Minnesota

11:15-11:40 Efficient Algorithms and Theory for High-Dimensional Bayesian Varying Coefficient Models

♦*Ray Bai*. University of South Carolina**Session 5E : Modern Streaming Data Analysis: Change-Point Problems And Applications**

Location: HPNP G114

Organizer: Jie Chen, Augusta University, Yajun Mei, Georgia Institute of Technology.

Chair: Ruizhi Zhang, University of Nebraska- Lincoln.

10:00-10:25 Detection of multiple change points in multiple profiles

♦*Jie Chen*¹ and ♦*Shirong Deng*². ¹Augusta University ²Wuhan University

10:25-10:50 Change-point Analysis of Hourly Sky-cloudiness Conditions in Canada

♦*Mo Li*¹, ♦*Qiqi Lu*¹ and ♦*Xiaolan Wang*². ¹Virginia Commonwealth University ²Environment and Climate Change Canada

10:50-11:15 Learning under concept drift

♦*Yuekai Sun*. University of Michigan

11:15-11:40 Inference for Gaussian Multiple Change-point Model via Bayesian Information Criterion
♦ *Yue Niu*¹, *Ning Hao*¹ and *Han Xiao*². ¹University of Arizona ²Rutgers University

Session 5F : Emerging Development In The Analysis Of Data With Complex Features

Location: HPNP G301

Organizer: Wenqing He, University of Western Ontario.

Chair: Wenqing He, University of Western Ontario.

10:00-10:25 Feature Screening with Large Scale and High Dimensional Survival Data

*Grace Yi*¹, ♦ *Wenqing He*¹ and *Raymond Carroll*².
¹University of Western Ontario ²Texas A&M University, University of Technology Sydney

10:25-10:50 Analysis of the Cox Model with Longitudinal Covariates with Measurement Errors and Partly Interval Censored Failure Times, with Application to an AIDS Clinical Trial

♦ *Yanqing Sun*¹, *Qingning Zhou*¹ and *Peter Gilbert*².
¹University of North Carolina at Charlotte ²Fred Hutchinson Cancer Research Center and University of Washington

10:50-11:15 Learning Optimal Dynamic Treatment Regimens Subject to Stagewise Risk Control

*Mochuan Liu*¹, *Yuanjia Wang*², *Haoda Fu*³ and ♦ *Donglin Zeng*¹. ¹University of North Carolina ²Columbia University ³Eli Lilly and Company

11:15-11:40 A new Bayesian method for handling covariate measurement error and detection limit in regression models

♦ *Muhire Kwizera*¹, *Roderick Little*², *Matthew Perzanowski*³ and *Qixuan Chen*¹. ¹Department of Biostatistics, Columbia University ²Department of Biostatistics, University of Michigan ³Department of Environmental Health Sciences, Columbia University

Session 5G : Statistical Leadership In Drug Development In The New Era Of Data Science

Location: HPNP 1101

Organizer: Yijie Zhou, Vertex Pharmaceuticals, Jun Zhao, Astellas Pharma.

Chair: Jun Zhao, Astellas Pharma.

10:00-10:25 Opportunities and Challenges of Using Real-world Data for Signal Identification and Evidence Generation to Inform Study Design and Scientific Questions in Medical Research
Yiyue Lou. Vertex Pharmaceuticals

10:25-10:50 Empowering Real-World Evidence Generation in Rare Conditions: Collaborative data initiatives

♦ *Jia Zhong*, *James Signorovitch* and *Eric Wu*. Analysis Group

10:50-11:15 Assessing Mediation Processes using Joint Longitudinal Models in the Framework of Individual Measurement Occasions

♦ *Jin Liu*¹, *Robert Perera*² and *Yijie Zhou*¹. ¹Vertex Pharmaceuticals ²Virginia Commonwealth University

11:15-11:40 Discussion

Yijie Zhou. Vertex

Session 5H: Student Paper Competition Winners

Location: HPNP 1102

Organizer: Organizing Committee.

Chair: Organizing Committee.

10:00-10:25 Sensitivity Analysis under the f-Sensitivity Models: A Distributionally Robust Optimization Viewpoint

♦ *Ying Jin*¹, *Zhimei Ren*² and *Zhengyuan Zhou*³. ¹Stanford University ²University of Chicago ³New York University

10:25-10:50 Fast Distributed Principal Component Analysis for Large-Scale Federated Data

♦ *Shuting Shen*, *Junwei Lu* and *Xihong Lin*. Harvard University

10:50-11:15 High-Dimensional Dynamic Process Monitoring By PCA-Based Sequential Learning

♦ *Xiulin Xie* and *Peihua Qiu*. University of Florida

11:15-11:40 Supervised Learning of Physical Activity Features from Functional Accelerometer Data

♦ *Margaret Banker* and *Peter X.K. Song*. University of Michigan

Special Invited Talks: Tue, June 21, 13:00-14:30 (EDT)

Session S1 : Special Invited Talks

Location: HPNP Auditorium (1404)

Organizer: Somnath Datta, Ph.D., University of Florida.

Chair: Ji-Hyun Lee, Ph.D., University of Florida.

13:00-13:45 Lessons Learned from the COVID-19 Pandemic: A Statistician's Reflection

Xihong Lin. Harvard University

13:45-14:30 Predictive model building through integration of information across disparate data sources and summary-statistics

Nilanjan Chatterjee. Johns Hopkins University

Sessions 6A-6H: Tue, June 21, 15:00-16:40 (EDT)

Session 6A : Recent Advances In Mendelian Randomization

Location: HPNP G312

Organizer: Chong Wu, Florida State University.

Chair: Chong Wu, Florida State University.

15:00-15:25 Inference of nonlinear causal effects with GWAS summary data

♦ *Ben Dai*¹, *Chunlin Li*², *Haoran Xue*², *Wei Pan*² and *Xiaotong Shen*². ¹The Chinese University of Hong Kong ²The University of Minnesota

15:25-15:50 Causal analysis with rerandomization estimators (CARE)

♦ *Chong Wu*¹ and *Jingshen Wang*². ¹FLORIDA STATE UNIVERSITY ²University of California, Berkeley

15:50-16:15 Breaking the Winner's Curse in Mendelian Randomization: Rerandomized Inverse Variance Weighted Estimator

*Xinwei Ma*¹, ♦ *Jingshen Wang*² and *Chong Wu*³. ¹UC San Diego ²UC Berkeley ³Florida State University

16:15-16:40 Constrained maximum likelihood-based Mendelian randomization robust to both correlated and uncorrelated pleiotropic effects

♦*Haoran Xue*¹, *Xiaotong Shen*² and *Wei Pan*¹. ¹Division of Biostatistics, School of Public Health, University of Minnesota ²School of Statistics, University of Minnesota

Session 6B : Recent Advances In Dimension Reduction Techniques

Location: HPNP G112

Organizer: Dipankar Bandyopadhyay, Virginia Commonwealth University.

Chair: Shanshan Ding, University of Delaware.

15:00-15:25 Significance testing for canonical correlation analysis in high dimensions

*Ian Mckeague*¹ and ♦*Xin Zhang*². ¹Columbia University ²Florida State University

15:25-15:50 Dimension Reduction Forests: Local Variable Importance using Structured Random Forests

♦*Joshua Loyal*¹, *Ruoqing Zhu*¹, *Yifan Cui*² and *Xin Zhang*³. ¹University of Illinois at Urbana-Champaign ²National University of Singapore ³Florida State University

15:50-16:15 Envelope model for function-on-function linear regression

♦*Zhihua Su*¹, *Bing Li*² and *Dennis Cook*³. ¹University of Florida ²Pennsylvania State University ³University of Minnesota

16:15-16:40 Floor Discussion.

Session 6C : Statistical Methods For Assessing Genomic Heterogeneity

Location: HPNP G101

Organizer: Yuchao Jiang, University of North Carolina at Chapel Hill.

Chair: Yuchao Jiang, University of North Carolina at Chapel Hill.

15:00-15:25 Robust Statistical Inference for Cell Type Deconvolution

♦*Jingshu Wang* and *Dongyue Xie*. University of Chicago

15:25-15:50 Single-cell eco-evolutionary dynamics of intratumor heterogeneity

Meghan Ferrall-Fairbanks. University of Florida

15:50-16:15 Neural Network Models for Sequence-Based TCR and HLA Association Prediction

♦*Si Liu*, *Phil Bradley* and *Wei Sun*. Fred Hutchinson Cancer Center

16:15-16:40 A statistical framework for cell-type-specific transcriptomics-wide association studies with an application to breast cancer

Xiaoyu Song. Icahn School of Medicine at Mount Sinai

Session 6D : Novel Statistical Modeling And Computing Methods For Complex Data

Location: HPNP G103

Organizer: Victor Hugo Lachos Davila, University of Connecticut.

Chair: Victor Hugo Lachos Davila, University of Connecticut.

15:00-15:25 New Bounded response models for target variables

Jorge Bazan. USP

15:25-15:50 Penalized complexity priors for the skewness parameter of power links

♦*Jose Ordonez*¹, *Marcos Prates*², *Jorge Bazan*³ and *Victor Lachos*⁴. ¹Federal University of Bahia ²Federal University of Minas Gerais ³ICMC - USP ⁴University of Connecticut

15:50-16:15 Linear Mixed-effects Models For Censored Data With Serial Correlation Errors Using The Multivariate Student's T-distribution

♦*Kelin Zhong*¹, *Rommy C. Olivari*², *Aldo M. Garay*² and *Victor H. Lachos*³. ¹Department of Statistics, UConn ²Department of Statistics, Federal University of Pernambuco ³Department of Statistics, University of Connecticut

16:15-16:40 Floor Discussion.

Session 6E : Modern Streaming Data Analysis: Detection And Identification

Location: HPNP G114

Organizer: Jie Chen, Augusta University, Ruizhi Zhang, University of Nebraska – Lincoln.

Chair: Ruizhai Zhang, University of Nebraska – Lincoln.

15:00-15:25 Low-Rank Robust Subspace Tensor Clustering for Metro Passenger Flow Modeling

Nurretin Sergin, *Jiuyun Hu* and ♦*Hao Yan*. Arizona State University

15:25-15:50 Optimal Parallel Sequential Change Detection under Generalized Performance Measures

*Zexian Lu*¹, *Yunxiao Chen*² and ♦*Xiaoou Li*¹. ¹University of Minnesota ²London School of Economics and Political Sciences

15:50-16:15 Differentially private approaches for streaming data analysis

Wanrong Zhang. Harvard University

16:15-16:40 Active sequential change-point detection under sampling control

Yajun Mei. Georgia Institute of Technology

Session 6F : Deep Learning With Application And Uncertainty Quantification

Location: HPNP G301

Organizer: Xinning Cui, University of California, Riverside.

Chair: Xinning Cui, University of California, Riverside.

15:00-15:25 Random walk with restart with graph embedded neural network to inform potential targets

♦*Yushi Liu*, *Bochao Jia* and *Rick Higgs*. Eli Lilly

15:25-15:50 Learning interactions in Reaction Diffusion Equation with Deep Learning

*Sichen Chen*¹, ♦*Nicolas Brunel*², *Xin Yang*³ and *Xinning Cui*¹. ¹Department of Statistics, University of California, Riverside ²Laboratoire de Mathématiques et Modélisation d'Evry, ENSIIE ³Department of Mathematics, University of California, Riverside

15:50-16:15 An optimal transport approach for selecting a representative subsample

Ping Ma. University of Georgia

16:15-16:40 Distribution-free uncertainty quantification for classification
♦ *Sasha Podkopaev and Aaditya Ramdas.* Carnegie Mellon University

Session 6G : Recent Advances In Clinical Trial Design And Practice

Location: HPNP 1101

Organizer: Shu Wang, University of Florida, Chung-Chou (Joyce) Chang, University of Pittsburgh.

Chair: Chung-Chou (Joyce) Chang, University of Pittsburgh.

15:00-15:25 A hybrid efficacy/effectiveness estimand for binary composite endpoints in clinical trials
♦ *Xingyuan Li and Nathan Morris.* Eli Lilly and Company

15:25-15:50 Bayesian adaptive model selection design for optimal biological dose finding in phase I/II clinical trials
Ruitao Lin. The University of Texas MD Anderson Cancer Center

15:50-16:15 A Simulation Study Evaluating Phase I Clinical Trial Designs for Combinational Agents
♦ *Shu Wang, Elias Sayour and Ji-Hyun Lee.* University of Florida

16:15-16:40 Bayesian Response Adaptive Randomization Design with A Composite Endpoint of Mortality and Morbidity
♦ *Zhongying Xu and Chung-Chou Chang.* University of Pittsburgh

Session 6H : New Developments In Modern Nonparametric Statistics And The Applications

Location: HPNP 1102

Organizer: Yichuan Zhao, Georgia State University.

Chair: Yichuan Zhao, Georgia State University.

15:00-15:25 Doubly robust U-statistic with applications
♦ *Ao Yuan, Anqi Yin and Ming Tan.* Georgetown University

15:25-15:50 Joint Semiparametric Models for Case-Cohort Designs
Weibin Zhong¹ and Guoqing Diao². ¹Bristol Myers Squibb ²George Washington University

15:50-16:15 Novel empirical likelihood inference for the mean difference with right-censored data
Kangni Alemjrodo¹ and Yichuan Zhao². ¹Purdue University ²Georgia State University

16:15-16:40 Asymptotic Normality of Gini Correlation in High Dimension with Applications to the K-sample Problem
♦ *Yongli Sang¹ and Xin Dang².* ¹University of Louisiana at Lafayette ²University of Mississippi

Sessions 7A-7H: Tue, June 21, 17:00-18:40 (EDT)

Session 7A : Novel Statistical Methods For -Omic Data Analysis

Location: HPNP G312

Organizer: Xiaoyu Song, Icahn School of Medicine at Mount Sinai.

Chair: Xiaoyu Song, Icahn School of Medicine at Mount Sinai.

17:00-17:25 LongStrain: An integrated strain-level analytic pipeline utilizing longitudinal metagenomics data
Boyan Zhou and Huilin Li. New York University

17:25-17:50 An all-in-one statistical framework that simulates realistic single-cell omics data and infers cell heterogeneity structure
Jingyi Jessica Li. UCLA

17:50-18:15 Data-Type Weighted Multi-Omics Spectral Clustering for Disease Subtyping
♦ *Peifeng Ruan and Hongyu Zhao.* Yale University

18:15-18:40 Deep Learning Methods for Retinal Imaging Genetics
Wei Chen. University of Pittsburgh

Session 7B : Modern Time Series And Network Methods In Data Science.

Location: HPNP G112

Organizer: Xinping Cui, University of California, Riverside, Ping Ma, University of Georgia.

Chair: Ping Ma, University of Georgia.

17:00-17:25 Collaborative Spectral Clustering in Attributed Networks
Pengsheng Ji. Univ. of Georgia

17:25-17:50 High Quantile Regression for Tail Dependent Time Series
Ting Zhang. University of Georgia

17:50-18:15 Dimension Reduction in Time Series Under the Presence of Conditional Heteroscedasticity
Murilo Dasilva, T. N. Sriram and Yuan Ke. University of Georgia

18:15-18:40 Multiple autocovariance changepoints problems in high-dimensional time series
Yuan Ke. University of Georgia

Session 7C : Innovative Approach Of Hidden Markov Model

Location: HPNP G101

Organizer: Hyoyoung Choo-Wosoba, National Cancer Institute, Center for Cancer Research, Biostatistics and Data Management Section.

Chair: Paul Albert, Branch Chief Senior Investigator.

17:00-17:25 Bayesian Semiparametric Hidden Markov Tensor Partition Models for Longitudinal Data with Local Variable Selection
Giorgio Paulon, Peter Mueller and Abhra Sarkar. UT-Austin

17:25-17:50 Non-Standard Applications of Hidden Markov Models in the Biosciences
♦ *Jordan Aron¹, Matthew O. Gribble², Li C. Cheung³, and Paul Albert³.* ¹University of Minnesota ²University of Alabama at Birmingham School of Public Health ³National Cancer Center

17:50-18:15 A hidden Markov model approach for a joinpoint trend analysis
♦ *Hyoyoung Choo-Wosoba, Philip Rosenburg and Paul Albert.* National Cancer Institute

18:15-18:40 Discussant: Paul Albert.

Session 7D : Statistical Advances And Applications In Analyzing Large Scale & Multi-Omic Single-Cell Data

Location: HPNP G103

Organizer: Rhonda Bacher, University of Florida; Department of Biostatistics.

Chair: Rhonda Bacher, University of Florida; Department of Biostatistics.

17:00-17:25 iscTrack, a semi-supervised algorithm and interactive single-cell tool to track emerging transcriptional states in serial samples
Jiannong Li, Scott Cukras, Sathya Sriramareddy, Keiran Smalley, Xiaoqing Yu and ♦Ann Chen. Moffitt Cancer Center

17:25-17:50 Deep learning methods for cell type identification and gene expression imputation
Sijie Yao, Xiaoqing Yu and ♦Xuefeng Wang. Moffitt Cancer Center

17:50-18:15 Nonparametric Interrogation of Transcriptional Regulation in Single-Cell RNA and Chromatin Accessibility Multiomic Data
Yuchao Jiang. UNC Chapel Hill

18:15-18:40 A statistical framework for scRNA-seq data modeling: simulation and applications
♦Guoshuai Cai¹, Xizhi Luo¹, Fei Qin¹ and Feifei Xiao².
¹University of South Carolina ²University of Florida

Session 7E : Modern Streaming Data Analysis: Process Monitoring

Location: HPNP G114

Organizer: Jie Chen, Augusta University, Yajun Mei, Georgia Institute of Technology.

Chair: Yajun Mei, Georgia Institute of Technology.

17:00-17:25 Fault Classification for High-dimensional Data Streams: A Directional Diagnostic Framework Based on Multiple Hypothesis Testing
Dongdong Xiang. East China Normal University

17:25-17:50 Adversarially Robust Sequential Hypothesis Testing
Shuchen Cao¹, ♦Ruizhi Zhang¹ and Shaofeng Zou².
¹University of Nebraska-Lincoln ²University at Buffalo, The State University of New York

17:50-18:15 Recent advances in quality and industrial analytics
Fugee Tsung. HKUST

18:15-18:40 Asymptotic Optimality Theory for Active Quickest Detection with Two Affected Streams
Qunzhi Xu. Georgia Institute of Technology

Session 7F : Discriminant And Cluster Analysis For Complex Data

Location: HPNP G301

Organizer: Xin (Henry) Zhang, Florida State University.

Chair: Guanyu Hu, University of Missouri.

17:00-17:25 Conditional probability tensor decompositions for multivariate categorical response regression
♦Aaron Molstad¹ and Xin Zhang². ¹University of Florida ²Florida State University

17:25-17:50 Quadratic Discriminant Analysis by Projection

Ruiyang Wu and ♦Ning Hao. University of Arizona

17:50-18:15 A Doubly-Enhanced EM Algorithm for Model-Based Tensor Clustering

♦Qing Mai, Xin Zhang, Yuqing Pan and Kai Deng. Florida State University

18:15-18:40 Stochastic Low-rank Tensor Bandits for Multi-dimensional Online Decision Making

Will Wei Sun. Purdue University**Session 7G : Design And Analysis In Vaccine Development And Its Challenges**

Location: HPNP 1101

Organizer: Bo Fu, Sanofi, Jun Zhao, Astellas.

Chair: Bo Fu, Sanofi.

17:00-17:25 Assessing the Role of Antibody in Vaccine Protection
Dean Follmann. NIH

17:25-17:50 Sensitivity Analysis for Evaluating Principal Surrogate Endpoints Relaxing the Equal Early Clinical Risk Assumption
♦Ying Huang, Yingying Zhuang and Peter Gilbert. Fred Hutchinson Cancer Research Center

17:50-18:15 Durability of Covid-19 Vaccines

Yu Gu. University of North Carolina

18:15-18:40 Statistical Consideration for Accelerated COVID-19 Vaccine Clinical Development in the Pandemic

James Zhou. HHS/ASPR/BARDA**Session 7H : Methods For Inference On Variable Importance Using Machine Learning (This session is co-sponsored by the Statistical Learning and Data Science (SLDS) Section of ASA)**

Location: HPNP 1102

Organizer: Brian Williamson, Kaiser Permanente Washington Health Research Institute.

Chair: Brian Williamson, Kaiser Permanente Washington Health Research Institute.

17:00-17:25 Inference for model-agnostic variable importance

♦Brian Williamson¹, Susan Shortreed¹, Peter Gilbert², Noah Simon³ and Marco Carone³. ¹Kaiser Permanente Washington Health Research Institute ²Fred Hutchinson Cancer Research Center ³University of Washington

17:25-17:50 Variable importance measure for spatial machine learning models with application to air pollution exposure prediction

♦Si Cheng, Ali Shojaie, Lianne Sheppard and Adam Szpiro. University of Washington

17:50-18:15 Floodgate: inference for model-free variable importance

♦Lu Zhang and Lucas Janson. Harvard University

18:15-18:40 Regularization on Ensembles of Tree and Variable importance

♦Siyu Zhou and Lucas Mentch. University of Pittsburgh

Banquet Talk: Tue, June 21, 20:00-20:45 (EDT) 10:00-10:25 Multilevel Modeling of Spatially Nested Functional Data: Spatiotemporal Patterns of Hospitalization Rates in the U.S. Dialysis Population

Session S2 : Banquet Talk

Location: Ben Hill Griffin Stadium Champions Club (121 Gale Lemerand Drive)

Organizer: ICSA Special Lecture Committee.

Chair: Samuel Wu, Ph.D., University of Florida.

20:00-20:45 Lost in translation

Lee-Jen Wei. Harvard University

Plenary Keynote Talk 3: Wed, June 22, 8:30-9:30 (EDT)

Session P3 : Plenary Keynote Talk 3

Location: HPNP Auditorium (1404)

Organizer: ICSA Special Lecture Committee.

Chair: Guogen Shan, Ph.D., University of Florida.

8:30-9:30 Inference for longitudinal data after adaptive sampling

Susan Murphy. Harvard University

Sessions 8A-8H: Wed, June 22, 10:00-11:40 (EDT)

Session 8A : Ultra-High Dimensional Variable Selection And Zero-Inflated Negative Binomial Spatial And Temporal Regression

Location: HPNP G312

Organizer: Hsin-Hsiung Huang, University of Central Florida.

Chair: Hsin-Hsiung Huang, University of Central Florida.

10:00-10:25 Spatiotemporal Zero-Inflated Bayesian Negative Binomial Regression Using Nearest Neighbor Gaussian Process and Polya-Gamma Mixtures

♦ *Qing He and Hsin-Hsiung Huang.* University of Central Florida

10:25-10:50 An Exchangeable Prior on Partitions for Clustering

♦ *Charles Harrison, Qing He and Hsin-Hsiung Huang.* University of Central Florida

10:50-11:15 Multi-Omics Integrative Analysis for Incomplete Data Using Weighted p-value Adjustment Approaches

Wenda Zhang¹, Joshua Habiger², Hsin-Hsiung Huang³ and ♦Yen-Yi Ho¹. ¹University of South Carolina ²Oklahoma State University ³University of Central Florida

11:15-11:40 Sparse Bayesian Matrix-variate Regression with High-dimensional Binary Response Data

♦ *Hsin-Hsiung Huang¹ and Shao-Shuan Wang².* ¹University of Central Florida ²National Central University

Session 8B : Recent Developments In Functional Data Analysis

Location: HPNP G112

Organizer: Gang Li, UCLA.

Chair: Gang Li, UCLA.

10:25-10:50 Online Estimation for Functional Data

Fang Yao. Beijing University

10:50-11:15 Functional ANOVA for High-Dimensional Spectral Analysis

♦ *Robert Krafty¹, Marie Tuft², Fabio Ferrarelli³, Ori Rosen⁴ and Zeda Li⁵.* ¹Emory University ²Sandia National Laboratory ³University of Pittsburgh ⁴University of Texas ⁵Baruch College, City University of New York

11:15-11:40 Factor-augmented model for functional data

Yuan Gao¹, Han Lin Shang² and ♦Yanrong Yang¹. ¹The Australian National University ²Macquarie University

Session 8C : Recent Advances In Robust Statistical Models For Censored And Missing Data

Location: HPNP G101

Organizer: Vicror Hugo Lachos Davila, University of Connecticut.

Chair: Jorge Luis Bazan Guzman, University of Sao Paulo.

10:00-10:25 Censored autoregressive regression models with Student-t innovations

Katherine Andreina Loorvaleriano¹, ♦Fernanda Langschumacher², Christian Galarza³ and Larissa Avilamatos¹. ¹University of Campinas ²Ohio State University ³Escuela Superior Politécnica del Litoral

10:25-10:50 Lasso regularization for censored skew-t regression and high dimensional predictors

Victor Hugo Lachos. University of Connecticut

10:50-11:15 Extending multivariate Student's-t semiparametric mixed models for longitudinal data with censored responses and heavy tails

Thalita Mattos¹, Victor Hugo Lachos², Luis Mauricio Castro³ and ♦Larissa Matos¹. ¹Universidade Estadual de Campinas ²University of Connecticut ³Pontificia Universidad Católica de Chile

11:15-11:40 Floor Discussion.

Session 8D : Recent Advances In Latent Variable Analysis

Location: HPNP G103

Organizer: Gongjun Xu, University of Michigan.

Chair: Xiangbin Meng, Northeast Normal University.

10:00-10:25 A Note on Statistical Inference for Noisy Incomplete Binary Matrix

Gongjun Xu. University of Michigan

10:25-10:50 VEMIRT: A Variational EM Algorithm-based Shiny App for High-dimensional IRT Applications

♦ *Chun Wang¹, Gongjun Xu², Chenchen Ma², Ruoyi Zhu¹ and Jiaying Xiao¹.* ¹University of Washington ²University of Michigan

10:50-11:15 A random effect hidden Markov model for process data

Xueying Tang. University of Arizona

11:15-11:40 Tree-informed Bayesian multi-source domain adaptation
♦Zhenke Wu¹, Zehang Li², Irena Chen¹ and Mengbing Li¹.
¹University of Michigan, Ann Arbor ²University of California, Santa Cruz

Session 8E : New Advances In Microbiome Related Data Analysis

Location: HPNP G114

Organizer: Zhigang Li, University of Florida.

Chair: Zhigang Li, University of Florida.

10:00-10:25 LinDA: Linear Models for Differential Abundance Analysis of Microbiome Compositional Data
♦Jun Chen¹ and Xianyang Zhang². ¹Mayo Clinic ²Texas A&M University

10:25-10:50 Identifying Microbial Interaction Networks Based on Irregularly Spaced Longitudinal Data
♦Jiang Gui¹, Jie Zhou¹, Weston Viles² and Annie Hoen¹.
¹Dartmouth College ²University of Southern Maine

10:50-11:15 Synergy Regression of Microbiome and Metabolome Data
Yue Wang. Arizona State University

11:15-11:40 A Novel Causal Mediation Analysis Approach for Zero-Inflated Count Mediators
♦Meilin Jiang¹, Seonjoo Lee², A. James O'malley³, Yaakov Stern² and Zhigang Li¹. ¹University of Florida ²Columbia University ³Geisel School of Medicine at Dartmouth

Session 8F : Statistical Computation Of Big Data With Biomedical Applications

Location: HPNP G301

Organizer: Sharmistha Guha, Texas A&M University, Statistics.

Chair: Sharmistha Guha, Texas A&M University, Statistics.

10:00-10:25 Bayesian data compression
♦Rajarshi Guhaniyogi¹ and Aaron Scheffler². ¹Texas A & M University ²UC San Francisco

10:25-10:50 A 'Divide-and-Conquer' AECM Algorithm for Large non-Gaussian Longitudinal Data
♦Reuben Retnam¹, Sanvesh Srivastava² and Dipankar Bandyopadhyay¹. ¹Virginia Commonwealth University ²University of Iowa

10:50-11:15 Bayesian Generalized Sparse Symmetric Tensor-on-Vector Regression
♦Sharmistha Guha and Rajarshi Guhaniyogi. Texas A&M University

11:15-11:40 Ordinal Causal Discovery
♦Yang Ni and Bani Mallick. Texas A&M University

Session 8G : Recent Development In Innovative Clinical Trial Designs

Location: HPNP 1101

Organizer: Dong Xi, Gilead Sciences.

Chair: Jiarui Lu, Novartis.

10:00-10:25 Graphical representation of the Hochberg procedure and other equally weighted tests
♦Dong Xi¹ and Frank Bretz². ¹Gilead Sciences ²Novartis

10:25-10:50 A unified framework for weighted parametric group sequential design (WPGSD)
Keaven Anderson, Zifang Guo, Jing Zhao and ♦Linda Sun. Merck & Co., Inc.

10:50-11:15 Statistical Interactions in a Clinical Trial
Naitee Ting. Boehringer Ingelheim Pharmaceuticals, Inc.

11:15-11:40 Deep historical borrowing framework in confirmatory clinical trials with multiple endpoints
Tianyu Zhan. AbbVie

Session 8H : Some Popular Applications In Data Integration

Location: HPNP 1102

Organizer: Gauri Datta, University of Georgia and U.S. Census Bureau.

Chair: Gauri Datta, University of Georgia and U.S. Census Bureau.

10:00-10:25 Multivariate Global-Local Priors for Small Area Estimation
Tamal Ghosh¹, ♦Malay Ghosh² and Jerry Maples³.
¹Citibank, Tampa ²University of Florida ³United States Census Bureau

10:25-10:50 Pseudo-Bayesian Small Area Estimation
♦Juhjung Lee¹, Gauri Datta² and Jiacheng Li².
¹University of Florida ²University of Georgia

10:50-11:15 Incorporating heterogeneous offsets in hierarchical disease mapping
♦Emily Peterson and Lance Waller. Emory University

11:15-11:40 Floor Discussion.

Sessions 9A-9H: Wed, June 22, 13:00-14:40 (EDT)

Session 9A : Bayesian Calibration Of Computer Models

Location: HPNP G312

Organizer: Vojtech Kejzlar, Skidmore College, Department of Mathematics and Statistics.

Chair: Tapabrata Maiti, Michigan State University, Department of Statistics and Probability.

13:00-13:25 On estimating photometric redshift of galaxies by augmenting observation with simulation
Arindam Fadikar. Argonne National Laboratory

13:25-13:50 A theoretical framework of the scaled Gaussian stochastic process in prediction and calibration
♦Mengyang Gu¹, Fangzheng Xie² and Long Wang³.
¹University of California, Santa Barbara ²Indiana University Bloomington ³Johns Hopkins University

13:50-14:15 Bayesian Calibration and Model Mixing
Matthew Pratola. Dept. of Statistics, The Ohio State University

14:15-14:40 An efficient approach for computer model calibration with variational Bayesian inference
♦Vojtech Kejzlar¹ and Taps Maiti². ¹Skidmore College ²Michigan State University

Session 9B : Novel Developments For Functional Data Analysis

Location: HPNP G112

Organizer: Raymond Wong, Texas A&M University.

Chair: Muxuan Liang, University of Florida.

13:00-13:25 Adaptive Frequency Band Analysis for Functional Time Series

♦ *Pramita Bagchi*¹ and *Scott Bruce*². ¹George Mason University ²Texas A&M University

13:25-13:50 Sliced Elastic Distance for Climate Model Validation

*Robert Garrett*¹, ♦ *Trevor Harris*² and *Bo Li*¹. ¹University of Illinois at Urbana-Champaign ²Texas A&M University

13:50-14:15 MARGARITA: Marginal-Product Basis Representation for Multi-dimensional Functional Data Analysis

William Consagra, *Arun Venkataraman* and ♦ *Xing Qiu*. University of Rochester

14:15-14:40 Floor Discussion.

Session 9C : Statistical Methods For High Dimensional Microbiome Data

Location: HPNP G101

Organizer: Somnath Datta, University of Florida, Subha Guha, University of Florida.

Chair: Subha Guha, University of Florida.

13:00-13:25 What Can We Learn About the Bias of Microbiome Studies from Analyzing Data from Mock Communities

*Mo Li*¹, ♦ *Glen Satten*², *Ni Zhao*¹, *Angel Rivera*³ and *Robert Tyx*³. ¹Johns Hopkins University ²Emory University ³CDC

13:25-13:50 Nonparametric Bayesian approaches for identifying differentially abundant genera between multiple groups in microbiome data

Archie Sachdeva, *Somnath Datta* and ♦ *Subharup Guha*. University of Florida

13:50-14:15 Deep ensemble learning over the microbial phylogenetic tree (DeepEn-Phy)

♦ *Wodan Ling*¹, *Youran Qi*², *Xing Hua*¹ and *Michael Wu*¹. ¹Fred Hutchinson Cancer Center ²Amazon

14:15-14:40 IFAA: Robust association identification and Inference For Absolute Abundance in microbiome analyses

Zhigang Li. University of Florida**Session 9D : Recent Advancements In Statistical Data Integration**

Location: HPNP G103

Organizer: Jin Jin, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health.

Chair: Jin Jin, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health.

13:00-13:25 Meta Clustering for Collaborative Learning

♦ *Chenglong Ye*¹, *Reza Ghanadan*² and *Jie Ding*³. ¹University of Kentucky ²Google ³University of Minnesota

13:25-13:50 Joint integrative analysis of dependent data sources

♦ *Emily Hector*¹ and *Peter Song*². ¹North Carolina State University ²University of Michigan

13:50-14:15 Data Integration Via Analysis of Subspaces

♦ *Jack Prothero*¹, *Meilei Jiang*², *Quoc Tran-Dinh*³, *Jan Hannig*³ and *J.s. Marron*³. ¹National Institute of Standards and Technology ²Meta ³UNC Chapel Hill

14:15-14:40 Synthetic-data-based transfer learning approach for multi-site risk prediction

♦ *Tian Gu* and *Rui Duan*. Department of Biostatistics, Harvard T.H. Chan School of Public Health**Session 9E : Modern Business Statistical Analysis**

Location: HPNP G114

Organizer: Aidong Adam Ding, Northeastern University, Shaobo Li, The University of Kansas.

Chair: Shaobo Li, The University of Kansas.

13:00-13:25 Penalized quantile regression

♦ *Ben Sherwood* and *Shaobo Li*. University of Kansas

13:25-13:50 On the use of Minimum Penalties in Multivariate Regression

♦ *Brad Price*¹ and *Ben Sherwood*². ¹West Virginia University ²University of Kansas

13:50-14:15 Joint Modeling of Playing Time and Purchase Propensity in Massively Multiplayer Online Role Playing Games Using Crossed Random Effects

Trambak Banerjee. University of Kansas

14:15-14:40 Measuring goodness-of-fit for bankruptcy prediction and its application to U.S. and Polish data

♦ *Xiaorui Zhu* and *Dungang Liu*. University of Cincinnati**Session 9F : Application And Theory Of Statistical Test And Evaluation**

Location: HPNP G301

Organizer: Aidong Adam Ding, Northeastern University.

Chair: Aidong Adam Ding, Northeastern University.

13:00-13:25 Statistical Evaluation of Deep Learning-based Side-channel Analysis

Aidong Ding. Northeastern University

13:25-13:50 Improved Meta-Analysis of ROC curves

♦ *Buddika Peiris*¹ and *Shuang Yang*². ¹Worcester Polytechnic Institute ²Worcester Polytechnic Institute

13:50-14:15 Signal-noise ratio of genetic associations and statistical power of SNP-set tests

*Hong Zhang*¹, *Ming Liu*², *Jiashun Jin*³ and ♦ *Zheyang Wu*². ¹Merck Research Laboratories ²WPI ³Carnegie Mellon University

14:15-14:40 BEAUTY Powered BEAST

*Kai Zhang*¹, ♦ *Zhigen Zhao*² and *Wen Zhou*³. ¹UNC ²Temple ³Colorado State University**Session 9G : Statistical Challenges In Clinical Trials For Alzheimer Disease**

Location: HPNP 1101

Organizer: Guoqiao Wang, Washington University in St. Louis, Changxing Ma, University at Buffalo.

Chair: Yan Li, Washington University in St. Louis.

13:00-13:25 Dose change and statistical power in the Aducanumab trial

Guogen Shan. University of Florida

- 13:25-13:50 A More Efficient Outcome for Alzheimer Disease Research: the Item Response Theory Based Score for the Clinical Dementia Rating (CDRR)
♦*Yan Li, Guoqiao Wang, Chengjie Xiong, Krista L Moulder and John C Morris.* Washington University in St. Louis
- 13:50-14:15 Floor Discussion.
- Session 9H : Statistics Education In The Era Of Ai And Data Science**
Location: HPNP 1102
Organizer: Steven Foti, University of Florida.
Chair: Steven Foti, University of Florida.
- 13:00-13:25 Interactive Graphics: A Bridge from Coding to Programming
Adam Loy. Carleton College
- 13:25-13:50 Case studies to community engagement: bringing hands-on data science experiences to the classroom
♦*Carrie Wright¹, Stephanie Hicks¹, Ava Hoffman¹, Michael Rosenblum¹, Michael Breshock¹, Qier Meng¹, Margaret Taub, Leah Jager¹, Tyler Derreth¹ and Mindi Levin¹.*
¹Johns Hopkins Bloomberg School of Public Health
- 13:50-14:15 Constructing a Modern Data Visualization Course: Topics, Reflections, and Feedback
Steven Foti. University of Florida
- 14:15-14:40 Foundations for NLP-assisted formative assessment feedback for short-answer tasks in large-enrollment classes
Susan Lloyd, ♦Matthew Beckman, Dennis Pearl, Rebecca Passoneau, Zhaohui Li and Zekun Wang. Penn State University

No.	Last Name	First Name	Title
1	Anyaso-Samuel	Samuel	Adjusting for informative cluster size in pseudo-value based regression approaches with clustered time to event data
2	Chakraborty	Nilanjana	A Bayesian framework for sparse estimation in High Dimensional Mixed Frequency Vector Autoregressive Models
3	Chan	Lap Sum	DrFARM: Identification and inference for pleiotropic variants in GWAS
4	Daw	Ranadeep	REDS: Random Ensemble Deep Spatial prediction
5	Dilma	Eleni	Class Distance model for community detection
6	Fang	Yusi	On p-value combination of independent and frequent signals: asymptotic efficiency and Fisher ensemble
7	Ge	Lin	Tailoring Capture-Recapture Methods to Estimate Registry-Based Case Counts Based on Error-Prone Diagnostic Signals
8	Hampton	Hayden	Deep Belief Network Anomaly Detection using Least Square Support Vector Methods
9	Han	Qiyu	Statistical Inference for Low Rank Matrix Regression with Adaptively Collected Data
10	Kang	Huining	A mixture model approach for identifying genes whose isoform abundances are associated with survival outcome
11	Kang	Tong	Analyzing Dental Fluorosis Data using a Novel Bayesian Model for Clustered Longitudinal Outcomes with an Inflated Category
12	Adhikary	Avizit Chandra	PARD: Patient-specific Abnormal Region Detection in Alzheimer's Disease Studies
13	Li	Wenhao	A Comparison of Two Approaches to Dynamic Prediction: Joint Modeling and Landmark Modeling
14	Lindberg	David	A Bayesian Nonparametric Approach to an HIV Assessment Survey with Missing Data and Skip Conditions
15	Liu	Jinyuan	A Distance-based Semiparametric Regression Framework for Between-subject attributes: Applications to High-dimensional Sequences of Microbiome and Wearables
16	Lu	Nicholas	Analyzing the Impact of Different Countries' Approaches to the COVID-19 Pandemic on Their Cumulative Infection Curves By Using Nonparametric Density Regression and Clustering Methods
17	Mao	Siqi	BERT based Financial Sentiment Index Enhanced (BERTFSIE) Models for Financial Markets Forecast
18	Yang	Yuting	Regression Analysis of a Future State Entry Time Distribution Conditional on a Past State Occupation in a Progressive Multistate Model
19	Roy	Samrat	A Regularized High Dimension Low Tubal-Rank Tensor Regression
20	Saha	Sudipto	A Constrained Bayesian Multiscale Spatial Model using the Truncated Normal Distribution
21	Samanta	Srijata	A generalized likelihood based Bayesian approach for scalable joint regression and covariance selection in high dimensions
22	Tan	Xiaoqing	Leveraging Models from Heterogeneous Data Sources to Improve Personalized Treatment Effect Estimation

23	Wang	Wei	Multivariate Survival Analysis in Big Data: A Divide-and-Combine Approach
24	Wang	Xing	Extreme and Inference for Tail Gini Functionals with Applications in Tail Analysis of Systemic Risk
25	Wang	Hongwei	Clinical Trials with External Control: Beyond Propensity Score Matching
26	Xie	Xiulin	Control Charts For Dynamic Process Monitoring With An Application To Air Pollution Surveillance
27	Yu	Mengxin	Are Latent Factor Regression and Sparse Regression Adequate?
28	Yue	Xiaowei	Physics-Constrained Bayesian Optimization
29	Zhang	Lu	StarTrek: Combinatorial Variable Selection with False Discovery Rate Control
30	Zhong	Weibin	Application of Two-step GONOGO Criteria and Model-based Design for Dose Finding Based on Efficacy
31	Zhou	Doudou	RISE: Rank in Similarity Graph Edge-Count Two-Sample Test
32	Zou	Jian	CGMM: an algorithm for constrained model-based clustering

Abstracts

Session 1A : Causal Inference And Its Applications

A causal approach to functional mediation analysis with application to a smoking cessation intervention

Donna Coffman

Temple University
dcoffman@temple.edu

The increase in the use of mobile and wearable devices now allow dense assessment of mediating processes over time. For example, a pharmacological intervention may have an effect on smoking cessation via reductions in momentary withdrawal symptoms. We define and identify the causal direct and indirect effects in terms of potential outcomes on the mean difference and odds ratio scales and present a method for estimating and testing the indirect effect of a randomized treatment on a distal binary variable as mediated by the nonparametric trajectory of an intensively measured longitudinal variable (e.g., from ecological momentary assessment). Coverage of a bootstrap test for the indirect effect is demonstrated via simulation. An empirical example is presented based on estimating later smoking abstinence from patterns of craving during smoking cessation treatment. We provide an R package, *funmediation*, to conveniently apply this technique. We conclude by discussing possible extensions to multiple mediators and directions for future research.

Estimating the Average Treatment Effect in Randomized Clinical Trials with All-or-None Compliance

Zhiwei Zhang

NIH/NCI
zhiwei.zhang@nih.gov

Noncompliance is a common intercurrent event in randomized clinical trials that raises important questions about analytical objectives and approaches. Motivated by the Multiple Risk Factor Intervention Trial (MRFIT), we consider how to estimate the average treatment effect (ATE) in randomized trials with all-or-none compliance. Confounding is a major challenge in estimating the ATE, and conventional methods for confounding adjustment typically require the assumption of no unmeasured confounders, which may be difficult to justify. Using randomized treatment assignment as an instrumental variable, the ATE can be identified in the presence of unmeasured confounders under suitable assumptions, including an assumption that limits the effect-modifying activities of unmeasured confounders. We describe and compare several estimation methods based on different modeling assumptions. Some of these methods are able to incorporate information from auxiliary covariates for improved efficiency without introducing bias. The different methods are compared in a simulation study and applied to the MRFIT.

Survey Weighting Strategies In Causal Mediation Analysis

Haoyu Zhou

Temple University
haoyu.zhou@temple.edu

Discussion: Causal Inference and its Applications

Esra Kurum

University of California, Riverside
esra.kurum@ucr.edu

Not applicable, discussant of the session.

Session 1B : Latent Variable Models In The Data Science Era

Identifiable Deep Generative Models via Sparse Decoding

♦Gemma Moran¹, Dhanya Sridhar², Yixin Wang³ and David Blei¹

¹Columbia University

²Mila and Universite de Montreal

³University of Michigan
gm2918@columbia.edu

We develop the sparse VAE for unsupervised representation learning on high-dimensional data. The sparse VAE learns a set of latent factors (representations) which summarize the associations in the observed data features. The underlying model is sparse in that each observed feature (i.e. each dimension of the data) depends on a small subset of the latent factors. As examples, in ratings data each movie is only described by a few genres; in text data each word is only applicable to a few topics; in genomics, each gene is active in only a few biological processes. We prove such sparse deep generative models are identifiable: with infinite data, the true model parameters can be learned. (In contrast, most deep generative models are not identifiable.) We empirically study the sparse VAE with both simulated and real data. We find that it recovers meaningful latent factors and has smaller heldout reconstruction error than related methods.

Population-Level Balance in Signed Networks

♦Weijing Tang and Ji Zhu

University of Michigan
weijtang@umich.edu

Statistical network models are useful for understanding the underlying formation mechanism and characteristics of complex networks. However, statistical models for signed networks have been largely unexplored. In signed networks, there exist both positive (e.g., like, trust) and negative (e.g., dislike, distrust) edges, which are commonly seen in real-world scenarios. The positive and negative edges in signed networks lead to unique structural patterns, which pose challenges for statistical modeling. In this paper, we introduce a statistically principled latent space approach for modeling signed networks and accommodating the well-known balance theory, i.e., "the enemy of my enemy is my friend" and "the friend of my friend is my friend". The proposed approach treats both edges and their signs as random variables, and characterizes the balance theory with a novel and natural notion of population-level balance. This approach guides us towards building a class of balanced inner-product models, and towards developing scalable algorithms via projected gradient descent to estimate the latent variables. We also establish non-asymptotic error rates for the estimates, which are further verified through simulation studies. In addition, we apply the proposed approach to an international relation network, which provides an informative and interpretable model-based visualization of countries during World War II.

Likelihood estimation of sparse topic distributions in topic models and its applications to Wasserstein document distance calculations

♦Xin Bing, Florentina Bunea, Marten Wegkamp and Seth-Strimas Mackey

Cornell University

xb43@cornell.edu

This paper studies the estimation of high-dimensional, discrete, possibly sparse, mixture models in topic models. The data consists of observed multinomial counts of p words across n independent documents. In topic models, the $p \times n$ expected word frequency matrix is assumed to be factorized as a $p \times K$ word-topic matrix A and a $K \times n$ topic-document matrix T . Since columns of both matrices represent conditional probabilities belonging to probability simplices, columns of A are viewed as p -dimensional mixture components that are common to all documents while columns of T are viewed as the K -dimensional mixture weights that are document specific and are allowed to be sparse. The main interest is to provide sharp, finite sample, l_1 -norm convergence rates for estimators of the mixture weights T when A is either known or unknown. For known A , we suggest MLE estimation of T . Our non-standard analysis of the MLE not only establishes its l_1 convergence rate, but reveals a remarkable property: the MLE, with no extra regularization, can be exactly sparse and contain the true zero pattern of T . We further show that the MLE is both minimax optimal and adaptive to the unknown sparsity in a large class of sparse topic distributions. When A is unknown, we estimate T by optimizing the likelihood function corresponding to a plug in, generic, estimator \hat{A} of A . For any estimator \hat{A} that satisfies carefully detailed conditions for proximity to A , the resulting estimator of T is shown to retain the properties established for the MLE. The ambient dimensions K and p are allowed to grow with the sample sizes. Our application is to the estimation of 1-Wasserstein distances between document generating distributions. We propose, estimate and analyze new 1-Wasserstein distances between two probabilistic document representations.

High-dimensional principle component analysis with heterogeneous missingness

◆ *Ziwei Zhu*¹, *Tengyao Wang*² and *Richard Samworth*³

¹University of Michigan, Ann Arbor

²London School of Economics

³University of Cambridge

ziweiz@umich.edu

We study the problem of high-dimensional Principal Component Analysis (PCA) with missing observations. Our main contribution is a new method, which we call primePCA, that is designed to cope with situations where observations may be missing in a heterogeneous manner. Given a good initialiser, primePCA iteratively projects the observed entries of the data matrix onto the column space of our current estimate to impute the missing entries, and then updates our estimate by computing the leading right singular space of the imputed data matrix. When the true principal components satisfy an incoherence condition and the signal is not too small, the error of primePCA provably converges to zero at a geometric rate. An important feature of our theoretical guarantees is that they depend on average, as opposed to worst-case, properties of the missingness mechanism. Our numerical studies on both simulated and real data reveal that primePCA exhibits very encouraging performance across a wide range of scenarios, including settings where the data are not Missing Completely At Random.

Session 1C : Some Advances In Statistical Machine Learning

Structurally Sparse Bayesian Neural Networks: Spike and Slab Shrinkage Priors

Sanket Jantre, ◆ *Shrijita Bhattacharya* and *Tapabrata Maiti*

Michigan State University

bhatta61@msu.edu

Network complexity and computational efficiency are increasingly significant aspects of deep learning. Sparse deep learning addresses these challenges by recovering the sparse structure of target functions while reducing over-parameterized model to a compact size. In this work, we adopt Bayesian solution through spike-and-slab group shrinkage priors to structurally reduce the network by pruning excess nodes. We propose variational Bayes inferences with continuous relaxation of discrete variables for posterior approximation.

An Adaptive Stochastic Approximation Algorithm for Randomized Decision GAN

Faming Liang

Purdue University

fmliang@purdue.edu

The Generative Adversarial Network (GAN) was recently introduced in the literature as a novel machine learning method for training generative models, which has many applications in statistics such as nonparametric conditional independence tests and nonparametric clustering. However, GAN is very difficult to train due to the issue of mode collapse, i.e., lack of diversity among generated data. We figure out the reason why GAN suffers from this issue and lay out a new theoretical framework for GAN based on randomized decision rules such that the mode collapse issue can be overcome essentially. Under the new theoretical framework, the discriminator converges to a fixed point while the generator converges to a distribution at the Nash equilibrium. We propose to train GAN by an adaptive stochastic gradient MCMC algorithm, where the generator is simulated from its conditional posterior distribution using a stochastic gradient MCMC algorithm, and the discriminator is treated as a hyperparameter of the posterior distribution of the generator and updated along with the simulation of the generator. The convergence of the proposed algorithm to the Nash equilibrium is asymptotically guaranteed. In addition, we show that the proposed algorithm is superior to classical ones in nonparametric conditional independence tests and nonparametric clustering. This talk is based on the joint work with Sehwan Kim and Qifan Song.

Volcano and valley prior with adhesive shrinkage for high dimensional data

Liangliang Zhang

case western reserve university

liangliangzhang.stat@gmail.com

The routine setups of Bayesian Variable selection assume that the prior of coefficient β is normal with mean 0 and variance σ^2 , and a hierarchical prior is used to control the scale of variance σ^2 . In contrast, we proposed a non-normal shrinkage prior with properties that the prior is symmetric which allows β to take positive and negative values, the prior has infinite density around 0 so that it will penalize true 0 coefficients to 0 and the prior has fat tails which give true nonzero coefficients freedom to be away from 0. The prior is controlled by two hyper-parameters. The shrinkage changes gradually when we modify the parameter values. We showed the effectiveness of our methods by comparing our model

with Horseshoe prior and Nonlocal prior. We discussed its connections with Bayesian discrete mixture priors. We also proved consistency theories and guided readers through real data applications.

Information-preserving Bayesian models for efficient and robust learning

Sandeep Madireddy

Argonne National Laboratory
smadireddy@anl.gov

Modern deep neural networks are in general quite brittle, and hence less robust to noise in the input or adversarial perturbation applied to them, as well as out-of-distribution data. From an information-theoretic point of view, the ability of the model to achieve better generalization and robustness will depend on whether the model learns more semantically meaningful information and compresses the nuisance information. Typically, the DNNs are trained by reducing an empirical loss function which can lead to nuisance or irrelevant information be memorized, that can have a detrimental impact on the model's robustness. To that end, we adopt an information-theoretic Bayesian approach that poses this intuition as a constrained optimization problem and hence seeks to learn a compressed (latent) encoding of input that is maximally informative about our target as measured through their mutual information. An important question is the choice of the latent dimension of the encoding so that the nuisance information is reduced. This question also appears in the context of popular generative models such as the variational autoencoders. To address this, we developed a Bayesian approach to model the joint distribution of the latent encoding dimension and the latent variable distribution through a spike and slab distribution and its variational formulation, which has the ability to systematically quantify the uncertainty across dimensions. The proposed approach shows good promise in terms of improving the model accuracy, robustness, and efficiency compared to traditional approaches deep learning approaches.

Session 1D : Machine Learning/Artificial Intelligence In Biomedical Research With 'big' Data

AI for Regulatory Science

Weida Tong

FDA
Weida.Tong@fda.hhs.gov

Artificial intelligence (AI) has made a significant mark in the past decade and demonstrated its utility in the broad area. The rapid advancement in AI also presents several opportunities and challenges to regulatory agencies with questions such as (1) how to assess and evaluate AI-based products and (2) how to develop and implement AI-based application to improve the agencies functions. In this presentation, the current thinking and on-going efforts at FDA in applying AI for risk and safety assessment will be discussed with a focus on "replace, reduce and refine" animal models with AI. AI consists of two application categories, predictive and generative; both are critical to predictive toxicology. Predictive algorithms learn from existing data/information to predict future outcomes, while generative algorithms produce new data with AI-driven study design. Examples will be given from the FDA projects for the latter. Specifically, testing data from laboratory animals provide crucial evidence for safety evaluation of therapeutics and risk assessment of environmental chemicals. This animal testing paradigm has been an essential component in regulating drug, food, and chemical safety by almost all the regulatory agencies in the world. As a result, abundant

animal data are available from public domain and private practice. As the toxicology community and regulatory agencies are moving towards reduction, refinement and replacement (3Rs principle) of animal studies, we are exploring an Artificial Intelligence (AI) approach to learn from the existing animal studies so that it can generate the animal data without conducting animal experiments. In this presentation, we introduce two AI models, AnimalGAN and ToxGAN, both were developed based on a generative adversarial network (GAN)-based framework. AnimalGAN was constructed to generate hematologic and clinical chemistry data while ToxGAN was for toxicogenomics data. Both models were capable of generating animal data only based on chemical information and experimental conditions (i.e., compound/time/dose combination). These results demonstrated the potential of utilizing the advanced AI approaches to produce non-animal models as alternative to animal study based on the existing data.

Causal networks for drug discovery

♦*Tao Xu*¹, *Shicheng Guo*², *Jinyung Zhao*¹ and *Momiao Xiong*³

¹University of Florida

²University of Wisconsin-Madison

³University of Texas Health Science Center at Houston
taoxu@ufl.edu

The classical paradigm for drug discovery is inefficient. Drug targets are often poorly defined for launched drug discovery and for potential therapeutic agents. We propose to change drug discovery paradigm from single drug target to multiple drug targets, from symptom to mechanism, from association analysis to cause discovery and from phenotype analysis to multiomics- based network pharmacology. The new paradigm for drug discovery includes (1) reconstruction of large-scale causal networks that offer a suitable framework to describe disease phenotypes and predict potential drug targets and (2) development of graphic neural networkbased perturbation analysis from the causal biological networks for drug target prediction. Learning causal networks from high dimensional data raises a great computational challenge. The classical methods for learning directed acyclic graphs (DAGs) are often formulated as a discrete optimization problem where the search space of DAGs is combinatorial. Therefore, inferring DAGs is a NP hard problem. To overcome this limitation, some researches recently formulate DAG reconstruction problem into a continuous optimization problem to dramatically reduce computational cost. The existing methods for construction of causal networks often focus on endogenous variables and ignore exogenous variables (covariates). To overcome this limitation, we develop models including both endogenous variables and exogeneous variables. We conduct extensive simulations and analyze potential factors on computational efficiency and structural estimation accuracy, including graph structure, error magnitude, network dimensions and sample size, and conclude that these new approaches outperform existing methods. The proposed algorithm has been applied to NIH's drug target discovery datasets including Riba virin, chloroquine and ruxolitinib.

New Toolkits for Disease Network Biology

Jake Chen

UAB Informatics Institute
jakechen@uab.edu

In this talk, I will first review the challenges in studying disease biology towards novel diagnostic and treatment insights. Multi-modal omics data have become widely available. However, it remains challenging to sift through cancer big data to prioritize genes, pathways, and gene signatures of interest for downstream translational

biomedical applications. We introduce four software tools from our lab to make disease network biology analysis easier, i.e., the Human Annotated Protein-Protein Interaction (HAPPI) database, the WINNER algorithm to rank genes in a connected disease biomolecular network, the WIPER algorithm to rank highly biologically significant protein-protein interactions in the same network, the BEERE software to identify and validate genes-phenotype associations in literature, the SEAS software to divide up patient samples into clinically enriched cohort groups, and the GeneTerrain software to characterize network gene signatures from functional genomic data sets. Overall, these tools provide new opportunities for both informaticians and biologists to address challenging big data problems in the future.

Achieving Differential Privacy with Matrix Masking in Big Data

Aidong Ding¹, ♦ Samuel Wu², Guanhong Miao² and Shigang Chen²

¹Northeastern University

²University of Florida
sw45@ufl.edu

Differential privacy schemes have been widely adopted in recent years to address issues of data privacy protection. We propose a new Gaussian scheme combining with another data protection technique, called random orthogonal matrix masking, to achieve differential privacy (DP) more efficiently. We prove that the additional matrix masking significantly reduces the rate of noise variance required in the Gaussian scheme to achieve DP in big data setting. With much less noise added, the resulting differential privacy protected pseudo data sets allow much more accurate inferences, thus can significantly improve the scope of application for differential privacy.

Session 1E : Statistical Challenges And Advances In Complex Data Analysis

Nontraditional Statistical Methods based on Wasserstein Distances and Conformal Prediction Set

Xiaoming Huo

Georgia Institute of Technology
huo@gatech.edu

We review our recent work in fast computational methods in optimal transport, which relates to the Wasserstein distance, and another project on the computation of conformal prediction set. New results will be presented and discussed.

Bayesian Spatially Varying Weight Neural Networks with the Soft-Thresholded Gaussian Process Prior

Jian Kang

University of Michigan
jiankang@umich.edu

Deep neural networks (DNN) have been adopted in the scalar-on-image regression which predicts the outcome variable using image predictors. However, training DNN often requires a large sample size to achieve a good prediction accuracy and the model fitting results can be difficult to interpret. In this work, we construct a novel single-layer Bayesian neural network (BNN) with spatially varying weights for the scalar-on-image regression. Our goal is to select interpretable image regions and to achieve high prediction accuracy with limited training samples. We assign the soft-thresholded Gaussian process (STGP) prior to the spatially varying weights and develop an efficient posterior computation algorithm based on stochastic gradient Langevin dynamics (SGLD). The BNN-STGP provides large prior support for sparse, piecewise-smooth, and continuous

spatially varying weight functions, enabling efficient posterior inference on image region selection and automatically determining the network structures. We establish the posterior consistency of model parameters and selection consistency of image regions when the number of voxels/pixels grows much faster than the sample size. We compared our methods with state-of-the-art deep learning methods via analyses of multiple real data sets including the task fMRI data in the Adolescent Brain Cognitive Development (ABCD) study.

Some Recent Advances on the analysis of Interval-Censored Case-cohort Failure Time Data

(Tony) Jianguo Sun

University of Missouri
sunj@missouri.edu

Case-cohort studies are commonly performed with the aim of reducing the cost of collecting covariate information and many authors have investigated their analyses when right-censored failure time data are available. In practice, however, it often happens that only interval-censored data are available. In this talk, we will discuss a couple of recently developed methods for the analysis of interval-censored case-cohort data, which include right-censored data as a special case.

An Efficient Method for Clustering Multivariate Longitudinal Data

Junyi Zhou¹, ♦ Ying Zhang² and Wanzhu Tu³

¹Amgen Inc

²UNMC

³Indiana University
ying.zhang@unmc.edu

Longitudinal data clustering is a challenging task, especially with sparse and irregular observations. It lacks reliable methods in the literature that deal with clustering complicated longitudinal data, particularly with multiple longitudinal outcomes. In this manuscript, a new agglomerative hierarchical clustering method is developed in conjunction with B-spline curve fitting and construction of unique dissimilarity measure for differentiating longitudinal observations. In an extensive simulation study, the proposed method demonstrates its superior performance in clustering accuracy and numerical efficiency compared to the existing methods. Moreover, the method can be easily extended to multiple-outcome longitudinal data without too much cost in computation and shows its robust results against the complexity of underlying mixture of longitudinal data. Finally, the method is applied to a date set from the SPRINT Study for validating the intervention efficacy in a Systolic Blood Pressure Intervention Trial and to a 12-year multi-site observational study (PREDICT-HD) for identifying the disease progression patterns of Huntington's disease (HD).

Session 1F : Statistical Methods And Applications For Analyzing Real-World Data

WeightP2V: a flexible risk prediction framework with patient representation weighted by medical concepts

Jia Guo and ♦ Shuang Wang

Columbia University
sw2206@columbia.edu

Electronic health records (EHRs) provide opportunities for researchers to develop clinical diagnostic tools. Conventional machine learning methods have been used to develop those tools, usually one tool for one diagnosis and cannot utilize information efficiently in longitudinal patient records. We developed a risk

prediction framework, WeightP2V, weighted patients to vectors. WeightP2V takes advantage of numeric representations of medical records of a patient based on which a numeric vector for the patient can be calculated with a weighting mechanism, i.e., medical records that are more relevant to an outcome of interest are given higher weights. Patient vectors are then used to predict patients' risks of developing any health outcomes in a database. With extensive simulation studies and clinical applications to predict 1,193 diagnoses in the MIMIC-III database and 1,559 diagnoses in the EHR database of Columbia University Irving Medical Center, we demonstrated an improved prediction performance of WeightP2V over competing methods.

Efficient Algorithms and Implementation of a Semiparametric Joint Model for Longitudinal and Competing Risks Data: With Applications to Massive Biobank Data

Shanpeng Li¹, Ning Li¹, Hong Wang², Jin Zhou¹, Hua Zhou¹ and Gang Li¹

¹UCLA

²Central South University
vli@ucla.edu

Semiparametric joint models of longitudinal and competing risks data are computationally costly and their current implementations do not scale well to massive biobank data. This paper identifies and addresses some key computational barriers in a semiparametric joint model for longitudinal and competing risks survival data. By developing and implementing customized linear scan algorithms, we reduce the computational complexities from $O(n^2)$ or $O(n^3)$ to $O(n)$ in various steps including numerical integration, risk set calculation, and standard error estimation, where n is the number of subjects. Using both simulated and real world biobank data, we demonstrate that these linear scan algorithms can speed up the existing methods by a factor of up to hundreds of thousands when $n > 10^4$, often reducing the runtime from days to minutes. We have developed an R-package, FastJM, based on the proposed algorithms for joint modeling of longitudinal and competing risks time-to-event data and made it publicly available on the Comprehensive R Archive Network (CRAN) at <https://CRAN.R-project.org/package=FastJM>.

A statistical quality assessment method for longitudinal observations in electronic health record data with an application to the VA million veteran program

Hui Wang¹, Ilana Belitskaya-Levy¹, Fan Wu¹, Jennifer Lee², Mei-Chiung Shih¹, Philip Tsao² and Ying Lu

¹Department of Veterans Affairs, Palo Alto, CA, USA

²Stanford University
ylul@stanford.edu

In this talk, we describe an automated method for assessment of the plausibility of continuous variables collected in the electronic health record (EHR) data for real world evidence research use. The most widely used approach in quality assessment (QA) for continuous variables is to detect the implausible numbers using pre-specified thresholds. In augmentation to the thresholding method, we developed a score-based method that leverages the longitudinal characteristics of EHR data for detection of the observations inconsistent with the history of a patient. The method was applied to the height and weight data in the EHR from the Million Veteran Program Data from the Veteran's Healthcare Administration (VHA). A validation study was also conducted, in which the receiver operating characteristic (ROC) metrics of the developed method outperforms the widely used thresholding method. We demonstrated that differ-

ent quality assessment methods have a non-ignorable impact on the body mass index (BMI) classification calculated from height and weight data in the VHA's database. In conclusion, the score-based method enables automated and scaled detection of the problematic data points in health care big data while allowing the investigators to select the high-quality data based on their need. Leveraging the longitudinal characteristics in EHR will significantly improve the QA performance.

Session 1G : Recent Advances In Survival And Recurrent Events Analysis For Complex Data Structures

Structured variable selection in Cox model with time-dependent covariates

◆ Guanbo Wang¹, Yi Yang¹, Mirelle Schnitzer², Tom Chen³, Rui Wang³ and Robert Platt¹

¹McGill University

²University of Montreal

³Harvard University
guanbo.wang@mail.mcgill.ca

To incorporate a wide range of covariate structures/selection dependencies, Mairal (2010) et al. developed the overlapping group Lasso, which allows a variable to be included in multiple groups. It was developed and applied mainly in the context of engineering and machine learning, where the outcome is often continuous or binary. In medical research, one common outcome of interest is time-to-event. Therefore, to broaden the use of the overlapping group Lasso, in this work, we extend it to accommodate data consisting of survival outcomes and time-dependent covariates. We present the algorithm of the overlapping group Lasso in the context of survival data in a straightforward way, which avoids knowledge of graph models. Nevertheless, how to group variable in practice to respect various selection dependencies have not well-studied. In this work, we provide roadmaps for grouping structure identification for common types of selection dependencies that can be used in the overlapping group Lasso. In the simulation study, we design complex covariate structures and show how to use the roadmaps. Various metrics of the performance of the estimator are assessed, and compared to the standard Lasso penalization in time-dependent Cox.

Robust Estimation for Recurrent Event Analysis in the Presence of Informative Event Censoring

◆ Tom Chen¹, Rui Wang¹ and Victor Degruttola²

¹Harvard Pilgrim Health Care and Harvard Medical School

²Harvard School of Public Health
tomchen00@gmail.com

Motivated by an "evolving cluster randomized trial" design for HIV prevention, where transmission clusters centered on newly diagnosed HIV individuals are established over time through phylogenetic analyses, we develop an estimating procedure for the intervention effect on patterns of HIV transmission in terms of the cluster sizes over time of the "evolving rings". We view each contact linked to the index case as a recurrent event to the index case and estimate treatment effects based on marginal rate and mean functions. A difficulty that arises is informative censoring of these contacts, which equates to missing events within the recurrent event process. We account for this dependent censoring through the use of inverse probability censoring weights.

Variance Estimation for Cox Model When Using Propensity Score Weighting

◆ Di Shu¹, Jessica G Young², Sengwee Toh² and Rui Wang²

¹University of Pennsylvania

²Harvard University

Di.Shu@Penncmedicine.upenn.edu

Inverse probability weighted Cox models can be used to estimate marginal hazard ratios under different point treatments in observational studies. To obtain variance estimates, the robust sandwich variance estimator is often recommended to account for the induced correlation among weighted observations. However, this estimator does not incorporate the uncertainty in estimating the weights and tends to overestimate the variance, leading to inefficient inference. In this talk, I will first introduce a new variance estimator that combines the estimation procedures for the hazard ratio and weights using stacked estimating equations, with additional adjustments for the sum of terms that are not independently and identically distributed in a Cox partial likelihood score equation. Then, I will extend the proposed variance estimator to accommodate clustered data. Finally, I will present a simulation study that compares the finite sample performance of the proposed method with alternative methods, and illustrate these different variance methods in both independent and clustered data settings, using a bariatric surgery data set and a multiple readmission data set, respectively.

Statistical Analysis of Recurrent Events from Administrative Databases

Yi Xiong

Fred Hutchinson Cancer Center

xiongyihill@gmail.com

Administrative health data contain rich information for investigating health issues; however, many restrictions and regulations apply to their use. Administrative health databases are maintained to serve non-research purposes and only data for people who seek health services is accessible. In addition, administrative health databases evolve over time and the regulations about their access may change. Motivated by administrative records of emergency department (ED) visits by children and youths in Alberta, we propose statistical methods to address two issues: (i) to evaluate dynamic pattern and impacts with doubly-censored recurrent event data and (ii) to re-calibrate estimators developed based on truncated information by leveraging summary statistics from the population. These methods are justified theoretically and numerically using both simulation and the ED visits data. This is a joint work with Dr. Joan Hu (Simon Fraser University) and Dr. Rhonda Rosychuk (University of Alberta).

Session 1H : Statistical Inference For Two-Phase Studies With Outcome-Dependent Sampling

Robust methods for Two-Phase Studies under generalized linear models

♦ *Jacob Maronge*¹, *Jonathan Schildcrout*² and *Paul Rathouz*³

¹University of Texas MD Anderson Cancer Center

²Vanderbilt University Medical Center

³Dell Medical School at the University of Texas at Austin
jmmaronge@mdanderson.org

Outcome-dependent sampling (ODS) is commonly used in two-phase design settings to increase efficiency for the estimation of an exposure of interest when response information (and possibly adjuster covariates) is available, but the exposure is expensive and/or cumbersome to collect. These two-phase studies are conducted as follows: in Phase One the response and adjuster covariate information is collected on a cohort that is representative of the target

population, but the expensive exposure is missing. In Phase Two, using response information from Phase One, we construct a sampling plan which selectively oversamples a subset of potentially informative subjects for which to collect expensive exposure information. There are two commonly-used likelihood-based methods for analyzing data from such studies, a conditional-likelihood approach and a full-likelihood approach. In essence, the full-likelihood retains incomplete Phase One data not selected into Phase Two. The resulting likelihood resembles those from missing data problems. In contrast, the conditional-likelihood explicitly conditions on Phase Two selection, not leveraging the additional information in leftover Phase One subjects. This results in a likelihood which reweights the Phase Two data by the sampling plan. In this talk, we contrast these methods using a novel semi-parametric extension to generalized linear models (SPGLM; Rathouz and Gao (2009)) to study and develop robust methods for analyzing data arising from these study designs when the response distribution may be misspecified. In this way, we aim to give practical design insights and flexible tools for practitioners in these settings.

Epidemiological Study Designs for Quantitative Longitudinal Data

♦ *Jonathan Schildcrout*, *Chiara Digraio* and *Ran Tao*

VUMC

jonny.schild@vumc.org

Outcome dependent sampling (ODS) designs are efficient when exposure ascertainment costs limit sample size. In longitudinal studies with a continuous outcome, ODS designs can lead to efficiency gains by using low-dimensional summaries of individual longitudinal trajectories to identify the subset of subjects in whom the expensive exposure variable will be collected. Analyses can be conducted using the outcome, confounder, and target exposure data from sampled subjects, or they can combine complete data from the sampled subjects with the partial data (i.e., outcome and confounder only) from the unsampled subjects. In this talk, we will discuss designs and analysis procedures for longitudinal and multivariate longitudinal data. We will examine finite sampling operating characteristics of design-analysis procedure combinations, and we will apply the various approaches to the analysis of lung function from subjects who participated in the Lung Health Study.

Statistical Methods for Selective Biomarker Testing in Two-Phase Studies

♦ *Natalie Delrocco*¹, *Adam Ding*² and *Samuel Wu*¹

¹University of Florida

²Northeastern University

ndelrocco95@ufl.edu

Studies investigating the association between clinical outcomes and relevant biomarkers are integral in many current statistical applications. When a large number of clinical outcomes are available, there are benefits to considering which units of observation are sampled for biomarker analysis, i.e. adopting a special case of outcome dependent sampling (ODS) where only those units with the most extreme clinical outcomes are biomarker typed. In this work, we employ a joint Gaussian assumption to derive point and interval estimates for the association between a continuous outcome and a biomarker under the above ODS sampling scheme. We show that this method is unbiased and more efficient than random sampling when assumptions are met. It is easy to implement using standard statistical software. The real-world performance is demonstrated in a chronic pain clinical trial.

Design and Analysis Strategies with "Secondary" Use Data*Sarah Lotspeich*

UNC

slotspeich@unc.edu

The growing availability of observational databases like electronic health records (EHR) provides unprecedented opportunities for secondary use of such data in biomedical research. However, these data can be error-prone and need to be validated before use. It is usually unrealistic to validate the whole database due to resource constraints. A cost-effective alternative is to implement a two-phase design that validates a subset of patient records that are enriched for information about the research question of interest. In this talk, I will discuss proper statistical approaches to analyze such two-phase studies, which can efficiently use the information in the unvalidated data in Phase I and address the potential biased validation sample selection in Phase II. I will demonstrate the advantages of the proposed methods over existing ones through extensive simulations and an application to an ongoing HIV observational study.

Session 2B : Advanced Research In Bio-Molecular And Imaging Data By Our Young Researchers**Outcome-guided Bayesian Clustering for Disease Subtype Discovery Using High-dimensional Transcriptomic Data***Lingsong Meng and ♦Zhiguang Huo*

Department of Biostatistics, University of Florida

zhuo@ufl.edu

The discovery of disease subtypes is an essential step for developing precision medicine, and disease subtyping via omics data has become a popular approach. While promising, subtypes obtained from conventional approaches may not be necessarily associated with clinical outcomes. The collection of rich clinical data along with omics data has provided an unprecedented opportunity to facilitate the disease subtyping process and to discover clinically meaningful disease subtypes. Thus, we developed an outcome-guided Bayesian clustering (GuidedBayesianClustering) method to fully integrate the clinical data and the high-dimensional omics data. A Gaussian mixed model framework was applied to perform sample clustering; a spike-and-slab prior was utilized to perform gene selection; a mixture model prior was employed to incorporate the guidance from a clinical outcome variable; and a decision framework was adopted to infer the false discovery rate of the selected genes. We deployed conjugate priors to facilitate efficient Gibbs sampling. Our proposed full Bayesian method is capable of simultaneously (i) obtaining sample clustering (disease subtype discovery); (ii) performing feature selection (select genes related to the disease subtype); and (iii) utilizing clinical outcome variable to guide the disease subtype discovery. The superior performance of the Guided-BayesianClustering was demonstrated through simulations and applications of breast cancer expression data. An R package has been made publicly available on GitHub to improve the applicability of our method.

Double soft-thresholded multigroup model for vector-valued image regression with application to DTI imaging*♦Arkaprava Roy¹ and Zhou Lan²*¹University of Florida²Yale University

ark007@ufl.edu

In this paper, we develop a novel spatial variable selection method for scalar on vector-valued image regression in a multi-group setting. Here, 'vector-valued image' refers to the imaging datasets that

contain vector valued information at each pixel/voxel location, such as in RGB color images, multimodal medical images, DTI imaging etc. The focus of this work is to identify the spatial locations in the image having important effect on the scalar outcome measure. However, individual components in the vector-valued predictor can be highly correlated, making the estimation step very difficult. We thus develop a novel shrinkage prior by soft-thresholding the l_2 norm of a latent multivariate Gaussian process. It will allow us to identify the spatial locations having non-zero effect from at least one of its components. Furthermore, motivated by our clinical application, the regression effect is decomposed into shared and group-specific parts, where a double soft-thresholding based prior is introduced. For posterior inference, an efficient MCMC algorithm is developed. We establish posterior contraction rate for parameter estimation and consistency for variable selection of the proposed Bayesian model, assuming that the true regression coefficients are Holder smooth. Finally, we illustrate our proposed method in simulations and in an ADNI dataset for modeling MMSE score based on DTI based vector-valued imaging markers.

Joint analysis and visualization of DNA methylation and nucleosome occupancy in single-molecule and single-cell data*Rhonda Bacher*

University of Florida

rbacher@ufl.edu

Alterations in DNA methylation, including the co-occurrence of both hyper- and hypo-methylation of different regions of the genome, have been detected in nearly all cancer types. In addition, both cancer- and tissue-specific differences exist in nucleosome positioning and occupancy, as well as transcription factor binding activity, which together determine chromatin accessibility. Methods to simultaneously capture nucleosome occupancy and methylation states at single-molecule and single-cell resolution are now possible. We present an approach we call methylscaper, a framework to simultaneously visualize DNA methylation and nucleosome occupancy states from joint experiments. We demonstrate methylscaper's ability to reveal biological patterns on both single-molecule and single-cell datasets.

Unity in diversity: Commonalities in these three different data analytical techniques*Arkaprava Roy, Rhonda L Bacher, Zhiguang Huo and ♦Susmita Datta*

University of Florida

susmdatta@gmail.com

As a discussant I will be presenting the commonalities of the statistical methodologies in all the three different talks on high-dimensional data analysis. Introduce some of the newest techniques dealing with high dimensional data. Primary purpose is to showcase some of the young researchers of the Department of Biostatistics at the University of Florida involved in novel methodological statistical research, involving high dimensional data, generated from advanced bio-molecular and brain imaging platform. All the three talks in the session will cover different aspects of Bayesian and frequentist statistical learning, statistical inference procedures and functional regression techniques, uniquely developed for not only high dimensional data but also for the integration of multiple types of them. I will discuss some generalized approach for integrating the combined knowledge transferable to any of these data analyses methods.

Session 2C : Emerging Topics In Statistical Learning For Biomedical Data

A Semiparametric Approach to Developing Well-calibrated Models for Predicting Binary Outcomes

♦ *Yaqi Cao*¹, *Ying Yang*² and *Jinbo Chen*¹

¹University of Pennsylvania

²Tsinghua University

jinboche@penncmedicine.upenn.edu

For evaluating the value of candidate predictors added to an existing risk prediction model, frequently data is only available from a "study source population" that deviates from the target population for prediction. Standard methods can be applied to develop an updated model from such data that accommodates both candidate and standard predictors, but the performance of the model in the target population cannot be evaluated in the absence of data from the target population. Considering that the model with only standard predictors is validated, we develop a novel statistical method for fitting the updated model that is guaranteed to be well-calibrated in the target population of prediction. We develop large-sample theory for this method, and evaluate its performance through extensive simulation studies as well as an application to data extracted from Penn Medicine Biobank to develop an improved breast cancer risk prediction model.

How does data preprocessing impact statistical learning in microRNA studies?

Li-Xuan Qin

MSKCC

qinl@mskcc.org

Statistical learning methods are booming used for analyzing high-dimensional biomedical data. While some of these methods make explicit or implicit use of data preprocessing such as variable pre-screening and variable standardization, other methods leave it up to the user to choose. The predictors trained by these methods are evaluated often using cross-validation, without realizing its complication by data artifacts resulted from experimental handling and the use of data normalization in data preprocessing for removing these artifacts. Towards an overarching goal of enhancing study replicability and reproducibility, we examined the connection between data preprocessing and statistical learning for sample classification in microRNA studies, using realistically distributed and robustly distributed data that were generated by re-sampling. Our study draws attention to the intricacy of data preprocessing choices for reproducible statistical learning and calls for attention on their mindful use and faithful reporting in research dissemination.

A Bayesian Reinforcement Learning Approach for Optimizing Combination Antiretroviral Therapy in People with HIV

♦ *Yanxun Xu*¹, *Wei Jin*¹, *Yang Ni*² and *Leah Rubin*¹

¹Johns Hopkins University

²Texas A&M University

yanxun.xu@jhu.edu

Numerous adverse effects (e.g., depression) have been reported for combination antiretroviral therapy (cART) despite its remarkable success on viral suppression in people with HIV (PWH). To improve long-term health outcomes for PWH, there is an urgent need to design personalized optimal cART with the lowest risk of comorbidity in the emerging field of precision medicine for HIV. Large-scale HIV studies offer researchers unprecedented opportunities to optimize personalized cART in a data-driven manner. However, the large number of possible drug combinations for cART makes the

estimation of cART effects a high-dimensional combinatorial problem, imposing challenges in both statistical inference and sequential decision-making. We develop a two-step Bayesian decision framework for optimizing sequential cART assignments. In the first step, we propose a dynamic model for individuals' longitudinal observations using a multivariate Gaussian process. In the second step, we build a probabilistic generative model for cART assignments and design an uncertainty-penalized policy optimization using the uncertainty quantification from the first step. Applying the proposed method to a dataset from the Women's Interagency HIV Study, we demonstrate its clinical utility in assisting physicians to make effective treatment decisions, serving the purpose of both viral suppression and comorbidity risk reduction.

HID machine: A Random Forest-based High Order Interaction Discovery Method for High-Dimensional Genomic Data

♦ *Min Lu*, *Yifan Sha* and *Xi Chen*

University of Miami

m.lu6@miami.edu

Although many methods of two-way interaction detection are available for genomic studies, detecting high-order interactions remains a major challenge due to the large number of variables in high-dimensional data. We developed a random forest-based algorithm, high order interaction discovery (HID), to efficiently select interaction signals. The random forest framework allows our method to analyze different types of outcomes including categorical, continuous and survival outcomes. In our method, an initial variable selection phase utilizing pairwise minimal depth indices is applied to choose potential interactive features. For interaction selection, we proposed and evaluated two variable-importance measures, minimal depth interaction importance (MDII) and permutation-based interaction importance (PBII). The MDII serves as a fast filter for potential interaction candidates, and PBII is used to finalize the ranking of the interaction terms. Our findings based on various simulation studies revealed that HID exhibited good and consistent performance in ranking interactions for different types of outcomes.

Session 2D : Statistics In Biosciences (Sibs): Real World Challenges And Recent Methodological Developments

Multi-sample single-cell RNA-seq data analysis and visualization - methods, software, and benchmark

♦ *Hongkai Ji*¹, *Boyang Zhang*¹, *Wenpin Hou*¹, *Zhicheng Ji*², *Zeyu Chen*³, *E John Wherry*³ and *Stephanie Hicks*

¹Johns Hopkins Bloomberg School of Public Health

²Duke University School of Medicine

³University of Pennsylvania Perelman School of Medicine

hji@jhu.edu

Single-cell RNA-seq experiments with multiple samples are increasingly used to discover cell types and their molecular features that may influence samples' phenotype (e.g. disease). However, analyzing and visualizing the complex phenotype-feature association remains nontrivial. Here I will introduce a framework for identifying and visualizing phenotype-associated gene and cell type abundance features from such data. The framework consists of a tree-based correlation screen method TreeCorTreat for identifying differential cell type abundance and cell-type specific differential genes, and a differential trajectory method Lamian for identifying differential cell type and gene features along pseudotemporal trajectories. Both methods are illustrated using an analysis of COVID-19 single-cell RNA-seq data. The analyses underscore the importance

of developing software and benchmark datasets in addition to new statistical methods in order to deal with the analytical challenges in applications of new technologies.

An efficient segmentation algorithm to estimate sleep duration from actigraphy data

Jonggyu Baek¹, ♦Margaret Banker², Erica Jensen², Xichen She², Karen Peterson², Andrew Pitchford³ and ♦Peter Song

¹University of Massachusetts Medical School

²University of Michigan

³Iowa State University
pxsong@umich.edu

Sleep duration is a recognized determinant of mental health, obesity and cardiovascular disease, cognition, and memory across the lifespan. Due to convenience and cost, sleep duration is often measured through self-report; yet, self-reported sleep duration can be highly biased. Actigraphy is a viable alternative as an objective measure of sleep. To analyze this actigraphy data, various sleep evaluation algorithms have been developed using regression methods, with coefficients constructed on minute-by-minute data measured at a specific device placement (wrist or hip). Because activity counts per minute may be affected by various factors in the study (e.g., type of device, sampling frequencies), regression-based algorithms developed within specific populations may not be generalizable to wider use. To address these concerns, we propose a new learning method to obtain robust and consistent sleep duration estimates. First, we identify temporal segments via pruned dynamic programming; then, we develop a calling algorithm with individual-specific thresholds and capture sleep periods. Our proposed method is motivated by and demonstrated in the Multi-Ethnic Study of Atherosclerosis (MESA) Sleep study and the Early Life Exposure in Mexico to Environmental Toxicants (ELEMENT) study.

Semiparametric estimation for length-biased interval-censored data with a cure fraction

Pao-Sheng Shen¹, ♦Yingwei Peng², Hsin-Jen Chen³ and Chyong-Mei Chen³

¹Tunghai University

²Queen's University

³National Yang Ming Chiao Tung University
yingwei.peng@queensu.ca

We considered length-biased and interval-censored data with a cure fraction arising from an early-onset diabetes mellitus study and proposed a new method to analyze such data in this talk. The Cox proportional hazards model for the survival time of the uncured individuals and the logistic regression model for the probability of being cured are employed to model the data. We construct the full likelihood function by assuming a homogeneous Poisson process for the incidence of the initial event and obtain semiparametric maximum likelihood estimates of the regression parameters by employing the EM algorithm. The large sample properties of the estimates are established. The performance of the method is assessed by simulations. The proposed model and method are applied to the data from the early-onset diabetes mellitus study.

Session 2E : Some Recent Methods For Sequential Monitoring Of Complex Data

A Robust Dynamic Screening System By Estimation of the Longitudinal Data Distribution

♦Lu You¹ and Peihua Qiu²

¹University of South Florida

²University of Florida

Lu.You@epi.usf.edu

To online monitor the longitudinal performance of processes and give early signals to processes with irregular patterns, a series of dynamic screening systems (DySS) have been proposed in the literature. Existing DySS methods are all based on estimation of the in-control (IC) mean and variance of processes with a regular longitudinal pattern. In this paper, a new DySS method is suggested, which is based on estimation of the IC distribution of processes with a regular longitudinal pattern. Based on the estimated IC distribution, a statistical process control chart is constructed for sequentially detecting any distributional shifts in a longitudinal process. The suggested control chart is relatively simple to design and implement, and it is robust to the true IC distribution. Numerical examples show that it outperforms some representative existing DySS methods. These properties make it an ideal tool for dynamic screening applications, which is demonstrated by a real-data example.

Transparent Sequential Learning for Statistical Process Control

Peihua Qiu

Founding Chair, Department of Biostatistics

pqi@ufl.edu

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

Statistical Quality Control Using Image Intelligence: A Sparse Learning Approach

Yicheng Kang

Bentley University

YKANG@bentley.edu

Advances in image acquisition technology have made it convenient and economic to collect large amounts of image data. In manufacturing and service industries, images are increasingly used for quality control purposes because of their ability to quickly provide information about product geometry, surface defects, and nonconforming patterns. In production line monitoring, image data often take the form of image streams in the sense that images from the process are being collected over time. In such applications, a fundamental task is to properly analyze image data streams. This image monitor-

ing problem is challenging for several reasons. First, images often have complicated structures such as edges and singularities, which render many traditional smoothing methods inapplicable. Second, a typical grayscale image contains tens of thousands of pixels, so the data is high-dimensional. It has been shown in the statistical process control (SPC) literature that conventional multivariate control charts have limited power of detecting process shifts when the data dimension is high. In this paper, we propose to transform images using a two-dimensional wavelet basis and monitor the wavelet coefficients by sparse learning-based multivariate control charts. By adapting the sparse learning algorithm to our quality control problem, the proposed method is able to detect shifts in the wavelet coefficients in a timely fashion and simultaneously identify those shifted coefficients. Combining this feature with the localization property of the wavelet basis, our method also enables accurate diagnosis of faulty image regions. In addition, the proposed charting statistics have explicit formulas, so they are easy to compute. Theoretical justifications and numerical comparisons with an existing method show that our method works well in applications.

Adaptive Process Monitoring Using Covariate Information

♦ Kai Yang¹ and Peihua Qiu²

¹Medical College of Wisconsin

²University of Florida

yklmy1994121@ufl.edu

Statistical process control charts provide a powerful tool for monitoring production lines in manufacturing industries. They are also used widely in other applications, such as sequential monitoring of internet traffic flows, disease incidences, health care systems, and more. In practice, quality variables are often affected in a complex way by many covariates, such as material, labor, weather conditions, social and economic conditions, and so forth. Among all these covariates, some could be observed, some might be difficult to observe, and the others might even be difficult for us to notice their existence. Intuitively, an SPC chart could be improved by using helpful information in covariates. However, because of the complex relationship between the quality variables and the covariates, shifts in the quality variables could be due to certain covariates whose data cannot be collected. On the other hand, shifts in some observable covariates may not necessarily cause shifts in the quality variables. Thus, it is challenging to properly use covariate information for process monitoring in a general setting. In this talk, we will introduce a method to handle this problem. An effective exponentially weighted moving average chart is developed, in which its weighting parameter is chosen large if the related covariates included in the collected data tend to have a shift and small otherwise. Because the covariate information is used in the weighting parameter only, the chart is designed solely for detecting shifts in the quality variables, but it can react to a future shift in the quality variables quickly because the helpful covariate information has been used in its observation weighting mechanism. Extensive numerical studies show that this method is effective in many different cases.

Session 2F : Big Data, Machine Learning And Graphical Methods

A Latent State Space Model for Learning Brain Dynamics for Mental Disorders

Yuanjia Wang

Columbia University

yw2016@cumc.columbia.edu

Modern neuroimaging technologies have substantially advanced the measurement of brain activities. Electroencephalogram (EEG) as a non-invasive neuroimaging technique measures changes in electrical voltage on the scalp induced by cortical activities. With its high temporal resolution, EEG has emerged as an increasingly useful tool to study brain connectivity. Challenges with modeling EEG signals of complex brain activities include interactions among unknown sources, low signal-to-noise ratio and substantial between-subject heterogeneity. We propose a state space model that jointly analyzes multi-channel EEG signals and learns dynamics of different sources corresponding to brain cortical activities. Our model borrows strength from spatially correlated measurements and uses low-dimensional latent sources to explain all observed channels. The model can account for patient heterogeneity and quantify the effect of a subject's covariates on the latent space. The EM algorithm, Kalman filtering, and bootstrap resampling are used to fit the state space model and provide comparisons between groups. We apply the developed approach to a case-control study of subjects at risk of alcoholism.

Clinical practice management of primary open-angle glaucoma in the United States: An analysis of real-world evidence

Joseph Imperato¹, Kelly Zou², Jim Li² and ♦Tarek Hassan³

¹IQVIA

²Medical Analytics and Real-World Evidence, Viatrix Inc

³Global Therapeutic Area Lead, Ophthalmology, Viatrix Inc

tarek.hassan@viatrix.com

To investigate clinical management of primary open-angle glaucoma (POAG) in the US using real-world evidence; to examine healthcare resource utilization (HCRU), medication adherence/persistence, and procedure use. A cross-sectional, retrospective analysis of Optum's de-identified Market Clarity Dataset. Patients ≥ 18 years with POAG diagnosis, and continuous enrollment for 1-year pre- and post-index were eligible and categorized into 4 mutually exclusive cohorts: CH1, treated with antiglaucoma medication(s) only; CH2, underwent glaucoma procedure(s) only; CH3, treated with antiglaucoma medication(s) and underwent procedure(s); CH4, received no treatment for POAG. Adherence and persistence with antiglaucoma medications, and disease specific HCRU were analyzed. Pairwise two-sample comparisons and multivariate regressions were conducted. Examined 232,572 eligible patients (CH1=60,895; CH2=4330; CH3=6027; CH4=161,320). Prostaglandins were most prescribed antiglaucoma medications (CH1: 69.7%; CH3: 62.7%), of which latanoprost was most common (CH1: 51.3%; CH3: 46.1%). Disease-specific office visits occurred in 26.3%, 78.2%, 75.0%, 23.8%, and surgical services visits occurred in 3.8%, 36.3%, 42.5%, 3.3%, in CH1-CH4, respectively. Adherence was higher (medication possession ratio: 47.1% vs. 39.4%; $P < .0001$), and more patients remained persistent across 1-year post-index period in CH1 vs. CH3 (25.4% vs. 16.1%; $P < .0001$). Positive predictors of medication persistence included being female, ≥ 55 years, and history of dyslipidemia or thyroid disease (all $P < .0003$).

Bayesian Pyramids: Identifiable Multilayer Discrete Latent Structure Models for Discrete Data

♦ Yuqi Gu¹ and David Dunson²

¹Columbia University

²Duke University

yuqi.gu@columbia.edu

High dimensional categorical data are routinely collected in biomedical and social sciences. It is of great importance to build in-

interpretable parsimonious models that perform dimension reduction and uncover meaningful latent structures from such discrete data. Identifiability is a fundamental requirement for valid modeling and inference in such scenarios, yet is challenging to address when there are complex latent structures. In this article, we propose a class of identifiable multilayer discrete latent structure models for discrete data, termed Bayesian pyramids. We establish the identifiability of Bayesian pyramids by developing transparent conditions on the sparsity structure of the pyramid-shaped directed graph. The proposed identifiability conditions can ensure Bayesian posterior consistency under suitable priors. As an illustration, we consider the two-latent-layer model and propose a Bayesian shrinkage estimation approach. Simulation results for this model corroborate identifiability and estimability of the model parameters. Applications of the methodology to DNA nucleotide sequence data uncover useful discrete latent features that are highly predictive of sequence types. The proposed framework provides a recipe for interpretable unsupervised learning of discrete data, and can be a useful alternative to popular machine learning methods.

Role of AI/ML and Big Data Analytics in Drug and Digital Medicine Development

Peter Zhang

Otsuka Pharmaceuticals (US)

peter.zhang@otsuka-us.com

Identify key considerations for developing novel digital endpoints for use in medical product development; Describe approaches to advancing the use of novel digital endpoints for use in medical product development; Evaluate the value of novel digital endpoints to different stakeholders including patients, clinicians, researchers, medical product developers, regulators, and payers.

Session 2G : Recent Development In Survival Analysis In Clinical Trials

On the Use of Restricted Mean Survival Time in Time-to-Event Data Analysis

Lihui Zhao

Northwestern University

lihui.zhao@northwestern.edu

Standard methods of summarizing the between-group difference with respect to a time-to-event outcome are based on Kaplan-Meier curves, the logrank test and the Cox's proportional hazards model. However, when the proportional hazards assumption is violated, the logrank test may not have sufficient power to detect the difference between two event time distributions, and the resulting hazard ratio is estimating a parameter involving the censoring time distribution and is difficult to interpret as a treatment contrast. On the other hand, the restricted mean survival time (RMST) is an easily interpretable, clinically meaningful summary of the survival function in the presence of censoring. The RMST is the mean survival time of all subjects in the study population followed up to a time point t and can be estimated consistently by the area under the Kaplan-Meier curve over $[0, t]$. In this research, we discuss the use of RMST for quantifying the between-group difference and its extension to the more general setting of competing risks. The resulting estimates have meaningful clinical interpretation. The data from clinical studies will be used for illustration.

Bayesian inference for a principal stratum estimand on recurrent events truncated by death

♦Tianmeng Lyu, Björn Bornkamp, Guenther Mueller-Velten and

Heinz Schmidli

Novartis

tianmeng.lyu@novartis.com

Recurrent events are often important endpoints in randomized clinical trials. For example, the number of recurrent disease-related hospitalizations may be considered as a clinically meaningful endpoint in cardiovascular studies. In some settings, the recurrent event process may be terminated by an event such as death, which makes it more challenging to define and estimate a causal treatment effect on recurrent event endpoints. In this talk, we focus on the principal stratum estimand where the treatment effect of interest on recurrent events is defined among subjects who would be alive regardless of the assigned treatment. For the estimation of the principal stratum effect in randomized clinical trials, we propose a Bayesian approach based on a joint model of the recurrent event and death processes with a frailty term accounting for within-subject correlation. We also present Bayesian posterior check procedures for assessing the model fit. The proposed approaches are demonstrated in the randomized Phase III chronic heart failure trial PARAGON-HF (NCT01920711).

A MCP-Mod approach to designing and analyzing survival trials with potential non-proportional hazards

♦Xiaodong Luo, Yuan Sun and Zhixing Xu

Sanofi

Xiaodong.Luo@sanofi.com

Non-proportional hazards have been observed in many studies especially in immuno-oncology clinical trials. Traditional analysis using the combined approach with log-rank test as the significance test and Cox model for treatment effect estimation becomes questionable as this approach relies heavily on the proportional hazards assumption. Inspired by the MCP-Mod (multiple comparisons and modeling approach) that has been widely used in dose-finding studies, we propose a similar approach to handle non-proportional hazards. Using this approach, efficacy signal is first established by a max-combo test, after which hazard ratios across time will be estimated using a logically nested splines model. Simulations studies and real-data examples are used to illustrate the use of this approach.

From Logic-respecting Efficacy Estimands to Logic-ensuring Analysis Principle for Time-to-event Endpoint in Randomized Clinical Trials with Subgroups

Yi Liu¹, Miao Yang¹, Siyoen Kil², Jiang Li³, Shoubhik Mondal⁴, Hong Tian³, Liwei Wang, ♦Yue Shentu⁵ and Godwin Yung⁶

¹Nektar Therapeutics

²LSK

³Beigene

⁴AstraZeneca

⁵Daiichi Sankyo Inc.

⁶Genentech

yshentu@dsi.com

An important goal of precision medicine is to identify biomarkers that are predictive, and tailor the treatment according to the biomarker levels of individual patients. Differentiating prognostic vs predictive biomarkers impacts important decision makings for patients and treating physicians. Using hazard ratio (HR) can mistake a purely prognostic biomarker for a predictive one leading to a disheartening possibility of depriving patients with beneficial treatment as demonstrated in the OAK trial. This stems from the illogical issue of HR at population level where marginal HR can be larger than those in both subgroups. Instead of trying to circumvent this issue by discouraging comparisons between marginal and subgroup HRs, we propose to directly fix it by using alternative logic

respecting efficacy estimands such as ratio of medians, ratio and difference of restricted mean survival times and milestone probabilities. These measures are straightforward, easy to interpret and clinically meaningful. More importantly, they will guarantee agreement between marginal and subgroup efficacy and provide cohesive message around efficacy profile of the drug. A further step in order to ensure logical estimates when analyzing real clinical trial data is the application of Subgroup Mixable Estimation (SME) Principle. Instead of inappropriately combining relative efficacy in subgroups for the marginal efficacy, the SME principle advocates following probability nature step by step with the key to mix two subgroups within each treatment arm using population or pooled sample prevalence on the probability scale. Detailed steps are provided for the aforementioned logic respecting efficacy estimands using either parametric or non-parametric approaches. Simultaneous inference can be provided with proper multiplicity adjustment to facilitate joint decision making with user-friendly apps.

Session 2H : Challenges And Recent Developments In Multi-Outcome Analysis

Alternative multivariate endpoints and related statistical models for clinical trials in Alzheimer disease

Guoqiao Wang

Division of Biostatistics, Washington University in St Louis
guoqiao@wustl.edu

Backgrounds The primary endpoint in clinical trials for Alzheimer disease is either a single test such as CDR SB or a composite of multiple tests such as the DIAN-TU composite. A composite score is usually an average of the individual components. These endpoints are typically analyzed using the univariate mixed effects model for repeated measures (MMRM) with time as a categorical variable and the statistical inference is often based on the single, end-of-study visit. However, there are multiple limitations for these endpoints. For instance, more than 20% of the placebo patients demonstrated no progression or improvement during an 18-month follow-up in CDR SB and more than 30% of the participants dropped out early. These participants will either dilute the treatment effect comparison or not contribute to the treatment effect estimation under the MMRM analysis. On the other hand, the composite is often limited to the combination of multiple components of the same type such as continuous cognitive endpoints, and thus may not present a comprehensive evaluation of the treatment effect while accounting for the early dropouts. The goal of this study is to propose some alternative multivariate endpoints and some innovative statistical models to analyze these endpoints. **Methods** We proposed to a multivariate endpoint that combines a longitudinal endpoint and a survival endpoint (i.e., dropout) and analyze it using the shared parameter proportional joint model (pJM). This multivariate endpoint can evaluate the cognitive treatment effect (i.e., in CDR SB) and the survival treatment effect (i.e., difference in dropout rate/timing) simultaneously, thus maximize the contribution of the early dropout participants. For the composite, we propose to analyze the components simultaneously using the multivariate model (i.e., the proportional MMRM [pMMRM]) so that the contribution of each component will not be overwhelmed by the other like in the composite. Results The shared parameter proportional joint model (pJM) can lead to more than 20% gain in power compared to the modelling of a single longitudinal endpoint or survival endpoint. For a composite with 4 components, modeling the components directly using multivariate

pMMRM can lead to 14% gain in power compared to model the composite using univariate pMMRM. **Conclusion** Multivariate endpoints and multivariate models (pJM/pMMRM) provide an alternative, more efficient way (e.g., greater power) to comprehensively evaluate the treatment effect simultaneously (cognitive benefits and survival benefits) and should be explored in clinical trials with longitudinal data.

Joint multivariate copula-frailty modeling of multiple-type recurrent events and the terminal event

Menglu Liang and [♦]Ming Wang

Penn State College of Medicine
mwang@phs.psu.edu

In clinical and observation studies, multiple type of recurrent events are often encountered, and they are likely to be highly correlated with each other, for instance, heart failure, myocardial infarction, stroke or other cardiovascular diseases. During the follow-up, these recurrent events can be censored by a competing risk event (e.g., death), leading to dependent censoring. Joint modeling of these multivariate recurrent events with the terminal event simultaneously is still an understudied field that guarantees more attention. Joint frailty model is one of the widely used approaches by sharing the same frailty term for recurrent and terminal event processes; however, the assumed conditional independence could be violated in practice. To address these issues, we propose a novel joint multivariate copula-frailty approach to model multiple types of recurrent events and a terminal event by incorporating the correlation between recurrent events and also their dependence with the terminal event. The Bayesian technique based on Metropolis-Hastings and the Gibbs sampling algorithms is utilized for parameter estimation and inference. Extensive simulation studies are conducted to evaluate the efficiency, robustness, and predictive performance of our proposal, and show its superiority compared to with the other competitive alternatives. In the end, we apply our method into the Cardiovascular Health Study for illustration.

Knowledge-guided Bayesian Factor Analysis for Integrative Analysis of Multi-Omics Data

Qiyiwen Zhang, Changgee Chang and [♦]Qi Long

University of Pennsylvania
qlong@penncmedicine.upenn.edu

High-dimensional multi-omics data offer great promises in advancing precision medicine. For example, they can be used to identify disease subgroups and uncover important molecular signatures predictive of disease risk and potential therapeutic targets. However, they also present daunting analytical challenges due to their high-dimensionality and heterogeneity. Factor analysis is a popular approach for extracting low-dimensional features from high-dimensional data. In addition, the knowledge-guided learning approach that use biological information such as functional genomics has been shown to improve prediction and feature selection in prior works. In this paper, we propose a novel Bayesian factor analysis model for integrative analysis of multi-omics data. It can handle both continuous and discrete data in a unified modeling framework using poly-gamma latent variables, and incorporate biological graph through a new prior formulation that is more flexible than the existing prior specifications. Simulation studies and real data analysis are conducted to demonstrate that the superior performance of our proposed method over existing methods.

Synergistic Self-learning Approach to Establishing Individualized Treatment Rules from Multiple Benefit Outcomes in a Cal-

Calcium Supplementation Trial♦ *Yiwang Zhou¹ and Peter Song²*¹Department of Biostatistics, St. Jude Children's Research Hospital²Department of Biostatistics, University of Michigan

yiwang.zhou@stjude.org

Precision nutrition is an emerging research field in nutritional sciences. Being a major risk to children's neurobehavioral and cognitive development, excessive in utero exposure to lead for embryos would be detrimental if no intervention is in place. The calcium supplementation trial conducted by the ELEMENT team aims to study the effect of daily calcium supplement in reducing maternal lead exposure to infants during pregnancy. This article focuses on establishing an individualized treatment rule (ITR) that can guide pregnant women on taking daily calcium supplementation to maximize the reduction of maternal lead exposure to infants. In the analysis we present a novel method, termed synergistic self-learning (SS-learning), to address two major challenges in the derivation of ITR in the presence of multiple clinical outcomes, including heterogeneous multidimensional outcomes and complex missing data patterns. SS-learning can effectively synergize heterogeneous features of multiple training data sources in the derivation of the ITR. Applied SS-learning to the ELEMENT calcium supplementation trial, we identified several important biomarkers that can be used to form an ITR that would give a higher expected lead reduction should it be implemented to the whole study population. We also examined the sensitivity and stability of SS-learning by comprehensive simulation studies.

Session 3A : Recent Advances In Statistical Methods For Causal Inference And Personalized Medicine**Evaluating different methods for estimating optimal treatment based on observational data**♦ *Qian Xu, Qi Zheng and Maiying Kong*

University of Louisville

maiying.kong@louisville.edu

Observational studies differ from experimental studies in that assignment of subjects to treatments is not randomized but rather occurs due to natural mechanisms, which are usually hidden from the researchers. Yet objectives of the two studies are frequently the same: identify the treatment effect of some exposure on a population, and identify which sub-population may benefit from a certain treatment. Many statistical methods, in the context of randomized experiments, are developed to achieve these objectives. Yet in an observational study special care must be taken to avoid confounding bias in treatment effect estimates, particularly when the number of covariates is large. In this talk, we develop a statistical model and estimating procedure to identify the variables which impact the treatment heterogeneity, which further help to identify which sub-population is more beneficial from a certain treatment. The estimators of the parameters in the model are consistent when propensity score model is correctly specified. We carried out extensive simulation studies to examine the proposed methods, which showed the proposed method performed well in identifying the characteristics of the patients who could be more beneficial from a certain treatment. We applied the proposed method to study which group of patients with heavy alcohol use are more beneficial from the treatment of varenicline based on a dataset from a randomized experiment. We also applied the proposed method to study which group of patients more likely benefit from statin use in control of inflammation for COVID-19 patients based on an observational study.

Evidence factors from multiple, possibly invalid, instrumental variables♦ *Anqi Zhao¹, Youjin Lee², Dylan Small³ and Bikram Karmakar⁴*¹National University of Singapore²Brown University³University of Pennsylvania⁴University of Florida

youjin.lee@brown.edu

Instrumental variables have been widely used to estimate the causal effect of a treatment on an outcome in the presence of unmeasured confounders. When several instrumental variables are available and the instruments are subject to possible biases that do not completely overlap, a careful analysis based on these several instruments can produce orthogonal pieces of evidence (i.e., evidence factors) that would strengthen causal conclusions when combined. We develop several strategies, including stratification, to construct evidence factors from multiple candidate instrumental variables when invalid instruments may be present. Our proposed methods deliver nearly independent inferential results each from candidate instruments under the more liberally defined exclusion restriction than the previously proposed reinforced design. We apply our stratification method to evaluate the causal effect of malaria on stunting among children in Western Kenya using three nested instruments that are converted from a single ordinal variable. Our proposed stratification method is particularly useful when we have an ordinal instrument of which validity depends on different values of the instrument.

Estimation of marginal treatment effect on binary outcome with multiple robustness and information borrow from secondary outcomes♦ *Chixiang Chen¹, Shuo Chen¹, Qi Long², Sudeshna Das³ and Ming Wang⁴*¹University of Maryland, School of Medicine²University of Pennsylvania³Harvard Medical School⁴Pennsylvania State University

Chixiang.Chen@som.umaryland.edu

The detection of the causal treatment effect on the development of certain disease is one of the key interests in clinical trials and biomedical studies. However, no one cannot pay on the success of inferring causal relationship when the treatment assignment is imbalanced and confounded by the other mechanisms. Specifically, when the treatment assignment is not randomized, the conventional logistic regression may not be valid to elucidate any causal inference, and exactly capturing all confounders is extremely difficulty in large-scale observational studies. In this paper, we propose a multiply robust (MultiR) framework to estimate a valid causal effect given a binary outcome, where multiple propensity score models and conditional mean imputation models are allowed in the estimation procedure. Moreover, we propose an enhanced MultiR (eMultiR) estimate that reduces estimation variability of MultiR estimate by incorporating secondary outcomes that are highly associated with the primary binary outcome. The resulting estimate is less sensitive to model mis-specifications compared to the existing state-of-the-art estimations (e.g., doubly robust estimate) via both theoretical and numerical assessments. The utility of (e)MultiR estimation is illustrated in one real data application from the National Alzheimer's Coordinating Center (NACC) data with the objective to detect the causal effect of short term use of antihypertensive medications on the development of dementia.

Estimation Of Marginal Treatment Effect On Binary Outcome

With Multiple Robustness And Information Borrow From Secondary Outcomes

Xiasuan Cai¹, Xinru Wang¹, Justin Baker², Jukka-Pekka Onnela³ and [♦]Linda Valeri¹

¹Columbia University

²McLean Hospital

³Harvard University
lv2424@cumc.columbia.edu

Mobile technology (e.g., mobile phones and wearable devices) enables unprecedented continuous monitoring of an individual's behavior, social interactions, symptoms, and other health conditions, presenting an enormous opportunity for therapeutic advancements and scientific discoveries regarding the etiology of psychiatric illness. Continuous collection of mobile data results in the generation of a new type of data: entangled multivariate time series of outcome, exposure, and covariates. Missing data is a pervasive problem in biomedical and social science research, and the Ecological Momentary Assessment (EMA) using mobile devices in psychiatric research is no exception. However, the complex structure of multivariate time series introduces new challenges in handling missing data for proper causal inference. Data imputation is commonly recommended to enhance data utility and estimation efficiency. The majority of available imputation methods are either designed for longitudinal data with limited follow-up times or for stationary time series, which are incompatible with potentially non-stationary time series. In the field of psychiatry, non-stationary data are frequently encountered as symptoms and treatment regimens may experience dramatic changes over time. To address missing data in possibly non-stationary multivariate time series, we propose a novel multiple imputation strategy based on the state space model (SSMmp) and a more computationally efficient variant (SSMimpute). We demonstrate their advantages over other widely used missing data strategies by evaluating their theoretical properties and empirical performance in simulations of both stationary and non-stationary time series, subject to various missing mechanisms. We apply the SSM impute to investigate the association between social network size and negative mood using a multi-year observational smartphone study of bipolar patients, controlling for confounding variables.

Session 3B : New Advances In High-Dimensional Data Analysis

On Statistical Inference with High Dimensional Sparse CCA

[♦]Nilanjana Laha¹, Nathan Huey¹, Brent Coull² and Raharshi Mukherjee¹

¹Harvard University

²Harvard Mukherjee
nlaha@hsph.harvard.edu

Many modern biomedical applications study association between complex factors, which can be high-dimensional by nature, but the association itself can be captured via low-dimensional structures. Studies involving multiple biological factors such as genetic markers, gene expressions, and disease phenotypes are typical examples. The traditional first step in these analyses is the assessment of linear association, formally known as the sparse canonical correlation analysis (SCCA). While SCCA succeeds at enforcing low dimensional structure (sparsity) through the likes of l_1 -penalty, it loses its amenability to classical inference such as testing and confidence intervals via traditional methods. We appeal to the popular debiasing technology to provide such inferential guarantees to SCCA. We also construct methodology for systematic variable selection in

this context. In the process, we discover that the variable selection procedure transcends from being computationally easy, to NP hard (subject to some recently popularized conjectures), to information-theoretically impossible as the low-dimensional structure becomes more complex.

A Graphical Lasso model for Hermitian matrices to detect global time-lagged Teleconnections

[♦]Indranil Sahoo¹, Joseph Guinness² and Brian J. Reich³

¹Virginia Commonwealth University

²Cornell University

³North Carolina State University
sahooi@vcu.edu

Teleconnections refer to spatially and temporally connected large-scale anomalies that influence the variability of atmospheric phenomena. Since teleconnections influence the global climate system, it is important to understand the abnormal behavior and interactions of these phenomena and identify them accurately. In this paper, we provide a mathematical definition of teleconnections based on a spatio-temporal model using spherical needlet functions. Spherical needlets are exactly localized at several overlapping intervals corresponding to different frequencies in the frequency domain and form a tight frame. This ensures the perfect reconstruction property of an orthonormal basis. We also extend the famous graphical Lasso algorithm to incorporate Hermitian matrices and use it to estimate the inverse covariance matrix of needlet coefficients after projecting them onto the Fourier domain. The proposed method is demonstrated by simulation studies and detection of possible global teleconnections in the HadCM3 model output air temperature data.

Multilayer Adjusted Cluster Point Process Model: Application to Microbial Biofilm Image Data Analysis

[♦]Suman Majumder¹, Brent Coull¹, Jessica Markwelch², Floyd Dewhirst³, Jacqueline Starr⁴ and Kyu Ha Lee¹

¹Harvard T.H. Chan School of Public Health

²Marine Biological Laboratories

³Forsyth Institute

⁴Brigham and Women's Hospital
smajumder@hsph.harvard.edu

A common challenge in spatial statistics is to quantify the spatial distributions of clusters of objects. Clusters of similar or dissimilar objects are encountered in many fields, including field ecology, astronomy, and biomedical imaging. Especially challenging is to quantify spatial clustering when one or more entities cluster around a different entity in multiple layers. Such multi-entity and multilayered structures are observed, for example, in human dental plaque biofilm images, which exhibit multispecies structures in corn-cob-like arrangements. We propose a novel, fully Bayesian, multivariate spatial point process model to quantify corn-cob-like arrangements with parent-offspring statistical approaches. The proposed multilayer adjusted cluster point process (MACPP) model departs from commonly used approaches in that it exploits the locations of the central "parent" object in clusters and accounts for multilayered multivariate parent-offspring clustering. In simulated datasets, the MACPP outperforms the classical Neyman-Scott process model, a univariate model for modeling spatially clustered processes, by producing decisively more accurate and precise parameter estimates. We analyzed data from a human dental plaque biofilm image in which Streptococcus and Porphyromonas simultaneously cluster around Corynebacterium, and Pasteurellaceae clusters around Streptococcus. The proposed MACPP model successfully captured the parent-offspring structure for all the taxa involved.

Correlated Wishart Matrices Classification via an Expectation-Maximization Composite Likelihood-Based Algorithm

Zhou Lan

Yale University
zhou.lan@yale.edu

Positive-definite matrix-variate data is becoming popular in computer vision. The computer vision data descriptors in the form of Region Covariance Descriptors (RCD) are positive definite matrices, which extract the key features of the images. The RCDs are extensively used in image set classification. Some classification methods treating RCDs as Wishart distributed random matrices are being proposed. However, the majority of the current methods preclude the potential correlation among the RCDs caused by the so-called auxiliary information (e.g., subjects' ages and nose widths, etc). Modeling correlated Wishart matrices is difficult since the joint density function of correlated Wishart matrices is difficult to be obtained. In this paper, we propose an Expectation-Maximization composite likelihood-based algorithm of Wishart matrices to tackle this issue. Given the numerical studies based on the synthetic data and the real data (Chicago face data-set), our proposed algorithm performs better than the alternative methods which do not consider the correlation caused by the so-called auxiliary information. All these above demonstrate our algorithm's compelling potential in image set classification in the coming future.

Session 3C : Machine Learning And Deep Learning Methods For Complex And Big Data

Generative models for diabetic retinopathy

Lingsong Zhang

Purdue University
lingsong@purdue.edu

In this talk, we will study the generative models for diabetic retinopathy. One part focuses on the simulation of the vessels in the retina, and the other part focuses on the background. Comparisons of existing approach and our proposed two stage method will also be discussed. This will help to early detection of diabetes progression of eye diseases.

Divide and conquer approaches for nonparametric regression and variable selection

Sapuni Chandrasena and ♦Rong Liu

University of Toledo
Rong.Liu@utoledo.edu

The rapid emergence of massive data with increasing size requests new statistical methods, especially in the fields of nonparametric regression, which is flexible but usually computationally intensive. To overcome the limitations of computing and storage, various distributed frameworks for statistical estimation and inference have been proposed. We study the statistical efficiency and asymptotic properties of the spline-backfitted kernel estimation for nonparametric additive models using the divide-and-conquer (DAC) approach. We also provide variable selection method based on majority voting procedure. Simulation study strongly supports the asymptotic theory and shows that the DAC approach is much more computational expedient without losing much accuracy.

A Bayesian Semi-supervised Approach to Keyword Extraction with Only Positive and Unlabeled Data

Guanshen Wang¹, ♦Yichen Cheng², Yusen Xia², Qiang Lin³ and Xinlei Wang¹

¹Southern Methodist University

²Georgia State University

³University of Science and Technology of China

ycheng11@gsu.edu

In the big data era, people are blessed with a huge amount of information. However, the availability of information may also pose great challenges. One big challenge is how to extract useful yet succinct information in an automated fashion. As one of the first few efforts, keyword extraction methods summarize an article by identifying a list of keywords. Many existing keyword extraction methods focus on the unsupervised setting, with all keyword assumed unknown. In reality, a (small) subset of the keywords may be available for an article. To utilize such information, we propose a rigorous probabilistic model based on a semi-supervised setup. Our method incorporates the graph-based information of an article into a Bayesian framework via an informative prior so that our model facilitates formal statistical inference, which is often absent in the existing methods. To overcome the difficulty arising from high-dimensional posterior sampling, we develop two Markov chain Monte Carlo algorithms based on Gibbs samplers, and compare their performance using benchmark data. We employ a false discovery rate (FDR) based approach for selecting the number of keyword, while the existing methods use ad-hoc threshold values. Our numerical results show that the proposed method compared favorably with state-of-the-art methods for keyword extraction.

Deep learning approaches for predicting virus-host interactions and drug response

Zhongming Zhao

University of Texas Health Science Center at Houston

Zhongming.Zhao@uth.tmc.edu

Virus infection is commonly observed in nature. Recently, SARS-CoV-2 has caused a global pandemic which has infected nearly 500 million individuals (as of April 3, 2022). An effective and efficient detection of viruses in the host genomes, and tracking how they interact with the host genomes, are currently a main challenge. In this talk, I will first introduce our computational approaches to detect viruses and their integration sites in the host genomes from next generation sequencing data. Then, based on our recently developed virus integration site database (VISDB), we developed a deep learning method, DeepVISP, for virus site integration prediction and motif discovery. To study COVID-19, we developed a deep learning method, DrivAER: Identification of Driving transcriptional programs with AutoEncoder derived Relevance scores from single cell RNA sequencing (scRNA-seq) data. We applied DrivAER to COVID-19 scRNA-seq data, as well as integrative analysis of COVID-19 genome-wide association studies (GWAS) and transcriptome-wide association studies (TWAS). Our investigation identified several genes, regulatory factors and cellular trajectories that might be involved in COVID-19 disease severity. Finally, I will present a deep generative neural network for accurate drug response imputation. In this work, we developed a deep variational autoencoder (VAE) model to compress thousands of genes into latent vectors in a low-dimensional space. We demonstrated that these encoded latent vectors could accurately impute drug response, outperform standard signature-gene based approaches, and appropriately control the overfitting problem.

Session 3D : Advance In Statistical Methods For Complex Data

Predicting long-term breast cancer risk with mammogram imaging data

♦ *Shu Jiang¹, Jiguo Cao², Bernard Rosner³ and Graham Colditz¹*

¹Washington university school of medicine

²Simon fraser university

³Harvard School of medicine

jiang.shu@wustl.edu

Screening mammography aims to identify breast cancer early and secondarily measures breast density to classify women at higher or lower than average risk for future breast cancer in the general population. Our primary goal in this study is to extract mammogram-based features that augment the well-established breast cancer risk factors to improve prediction accuracy. In this talk, I will present a novel supervised functional principal component analysis to extract image-based features that are ordered by association with the failure times.

Fighting Noise with Noise: Causal Inference with Many Candidate Instruments

♦ *Xinyi Zhang, Linbo Wang, Stanislav Volgushev and Dehan Kong*

University of Toronto

xyi.zhang@mail.utoronto.ca

Instrumental variable methods provide useful tools for inferring causal effects in the presence of unmeasured confounding. To apply these methods with large-scale data sets, a major challenge is to find valid instruments from a possibly large candidate set. In practice, most of the candidate instruments are often not relevant for studying a particular exposure of interest. Moreover, not all relevant candidate instruments are valid as they may directly influence the outcome of interest. In this article, we propose a data-driven method for causal inference with many candidate instruments that addresses these two challenges simultaneously. A key component of our proposal is a novel resampling method that constructs pseudo variables to remove irrelevant candidate instruments having spurious correlations with the exposure. Theoretical and synthetic data analyses show that the proposed method performs favourably compared to existing methods. We apply our method to a Mendelian randomization study estimating the effect of obesity on health-related quality of life.

Smooth nonparametric dynamic prediction for competing risks via deep learning

Zhiyang Zhou

University of Manitoba

zhiyang.zhou@umanitoba.ca

In the risk prediction for medicine, public health, economics, engineering, and many other areas, one may have to handle competing risks (i.e., mutually exclusive events) and figure out the relationship between their incidence probabilities and risk factors. Although the recent success of risk prediction has already been extended to the dynamic version (where time-varying risk factors are incorporated into models), existing approaches usually involve strong assumptions (e.g., additive effects and/or proportional hazard) which may lead to extra bias in prediction. To tackle these issues, we present an output layer named the Smooth Monotonic Output Layer (SMOL). When concatenated to deep neural networks, SMOL may help us learn incidence probabilities directly without specifying a parametric structure. We conducted numerical experiments on data collected through the Lifetime Risk Pooling Project (LRPP) which pooled together twenty community-based studies on cardiovascular

disease — the leading cause of death in the world — and involved around three hundred thousand participants with long-term follow-ups of longitudinal risk factors. Extensive results showed a state-of-the-art accuracy of our proposal in predicting individual risks of cardiovascular diseases and non-cardiovascular death simultaneously.

Distributed Cox Proportional Hazards Model Using Summary-level Information

♦ *Dongdong Li¹, Wenbin Lu², Di Shu³, Sengwee Toh¹ and Rui Wang⁴*

¹Harvard Medical School

²North Carolina State University

³University of Pennsylvania Perelman School of Medicine

⁴Harvard Medical School and Harvard T.H. Chan School of Public Health

dongdong.li@hphci.harvard.edu

Individual-level data sharing across multiple sites can be infeasible due to privacy and logistical concerns. This work proposes a general distributed methodology to fit Cox proportional hazards models without sharing individual-level data in multi-site studies. We make inferences on the log hazard ratios based on an approximated partial likelihood score function that uses only summary-level statistics. This approach can be applied to both stratified and unstratified models, accommodate both discrete and continuous exposure variables, and permit the adjustment of multiple covariates. In particular, the fitting of stratified Cox models can be carried out with only one file transfer of summary-level information. We derive the asymptotic properties of the proposed estimators and compare the proposed estimators with the maximum partial likelihood estimators using pooled individual-level data and meta-analysis methods through simulation studies. We apply the proposed method to a real-world data set to examine the effect of sleeve gastrectomy versus Roux-en-Y gastric bypass on the time to first postoperative readmission.

Session 3E : Recent Advancement In Statistical Learning Methods For High-Dimensional Biomedical Data

On p-value combination of independent and frequent signals: asymptotic efficiency and Fisher ensemble

♦ *Yusi Fang¹, Chung Chang² and George Tseng¹*

¹Biostatistics, University of Pittsburgh

²Applied Math, National Sun Yat-sen University

ctseng@pitt.edu

Combining p -values to integrate multiple effects is of long-standing interest in social science and biomedical research. In this paper, we focus on revisiting a classical scenario closely related to meta-analysis, which combines a relatively small (finite and fixed) number of p -values while the sample size for generating each p -value is large (asymptotically goes to infinity). We evaluate a list of traditional and recently developed modified Fisher's methods to investigate their asymptotic efficiencies and finite-sample numerical performance. The result concludes that Fisher and adaptively weighted Fisher method have top performance and complementary advantages across different proportions of true signals. Consequently, we propose an ensemble method, namely Fisher ensemble, to combine the two top-performing Fisher-related methods using a robust truncated Cauchy ensemble approach. We show that Fisher ensemble achieves asymptotic Bahadur optimality and integrates the strengths of Fisher and adaptively weighted Fisher methods in simulations.

We subsequently extend Fisher ensemble to a variant with emphasized power for concordant effect size directions. A transcriptomic meta-analysis application confirms the theoretical and simulation conclusions, generates intriguing biomarker and pathway findings, and demonstrates the strengths and strategy of using the proposed Fisher ensemble methods.

Improve Health Equality for Polygenic Risk Score (PRS) by Joint Penalized Regression of GWAS Summary Statistics from Two Ancestries

♦ Peng Liu¹, Max G'sell¹, Bernie Delvin² and Kathryn Roeder¹

¹Carnegie Mellon University

²University of Pittsburgh
pengl2@andrew.cmu.edu

The polygenic risk scores (PRSs) are quantitative metrics for the genetic risk of disease. PRSs for schizophrenia are largely developed using the European population. Consequently, the scores suffer a significant loss in prediction accuracy when applied to subjects of non-European origin. To help mitigate the racial disparities in PRS of schizophrenia, we proposed Joint-Lassosum, a powerful and computational efficient PRS calculation method in this paper. Joint-Lassosum uses a joint integrative penalized regression model to leverage the genetic information from both European and non-European populations. From extensive simulations, we have shown that our method can improve the portability of PRS when compared with other methods. Another interesting finding is that the improved PRS might help with diagnosis bias in schizophrenia. A software implementation of our algorithm has also been developed.

High-dimension to high-dimension screening for detecting genome-wide epigenetic regulators of gene expression

Hongjie Ke¹, Zhao Ren², Shuo Chen¹, George Tseng², Jianfei Qi¹ and ♦Tianzhou Ma¹

¹University of Maryland

²University of Pittsburgh
tma0929@umd.edu

The advancement of high-throughput technology characterizes a wide range of epigenetic modifications across the genome involved in disease pathogenesis via regulating gene expression. The high-dimensionality of both epigenetic and gene expression data make it challenging to identify the important epigenetic regulators of genes. Conducting univariate test for each epigenetic-gene pair is subject to serious multiple comparison burden, and direct application of regularization methods to select epigenetic-gene pairs is computationally infeasible. Applying fast screening to reduce dimension first before regularization is more efficient and stable than applying regularization methods alone. We propose a novel screening method based on robust partial correlation to detect epigenetic regulators of gene expression over the whole genome, a problem that includes both high-dimensional predictors and high-dimensional responses. Compared to existing screening methods, our method is conceptually innovative that it reduces the dimension of both predictor and response, and screens at both node (epigenetic features or genes) and edge (epigenetic-gene pairs) levels. We develop data driven procedures to determine the conditional sets and the optimal screening threshold, and implement a fast iterative algorithm. Simulations and two applications to long non-coding RNA and DNA methylation regulation in Kidney cancer and Glioblastoma Multiforme illustrate the validity and advantage of our method.

The mediating role of neuroimaging data in age-related cognitive decline

♦Hwiyoung Lee and Shuo Chen

University of Maryland, Baltimore
hwiyoung.lee@som.umaryland.edu

Aging changes brain functions and structures in a downward trajectory and consequently causes decayed neurocognitive performance. To further understand the neurophysiology of this process, we investigate the mediation role of multivariate neuroimaging variables in age-related cognitive decline. Considering the fact that cognition is exclusively determined by the brain, we propose a new multivariate mediation model by maximizing the mediation effect of brain imaging data. Specifically, we decompose the total effect of aging on cognitive function into natural direct and indirect effects and maximize the indirect effect with a parsimonious set of neuroimaging variables. We implement the optimization problem by alternating direction method of multipliers, and consider the aggregate effect of selected imaging variables as a functional brain age score (FBAS). The simulation results show that our method can accurately select imaging variables and estimate the mediation effect in comparison to existing methods. We further apply the proposed method to the whole-brain cortical thickness and white-matter integrity measures of 37,441 UK Biobank participants, and find that the mediation effect of brain imaging variables explains more than 91% of the age-related cognitive decline.

Session 3F : Advanced Statistical Learning Methods For Dynamic Systems

A Computing Algorithm for Parameter Estimation of Ultra-high Dimensional VAR Model

Hongyu Miao

Florida State University
hmiao@fsu.edu

In this study we tackled the parameter estimation problem for dynamic models with millions of unknown parameters, using time-course sparse data. We proposed a novel formulation and an efficient model fitting algorithm based on the topological properties of the corresponding graphical model. Both simulation studies and real data application demonstrated the superiority of the proposed method.

Generalized Ordinary Differential Equation (GODE) Model and Its Link to Deep Learning

Hulin Wu

University of Texas Health Science Center at Houston
Hulin.Wu@uth.tmc.edu

We proposed a generalized ordinary differential equation (GODE) model (Miao, Wu and Xue, JASA, 2014) to quantify the dynamics of discrete data that follow an exponential distribution. Recently a connection between ordinary differential equation (ODE) models and deep learning models was established. In particular, deep residual networks (ResNets) was reparameterized and interpreted as solutions to a continuous ordinary differential equation or Neural-ODE model. In this study, we propose a neural generalized ordinary differential equation (Neural-GODE) model with layer-varying parameters to further extend the Neural-ODE to flexibly approximate the discrete ResNet. Specifically, we use non-parametric B-spline functions to parameterize the coefficients in the Neural-GODE so that the trade-off between the model complexity and computational efficiency can be easily balanced. It is demonstrated that the ResNets and Neural-ODE models are special cases of the proposed Neural-GODE model. Based on two benchmark datasets, MNIST and CIFAR-10, we show that the layer-varying

Neural-GODE is more flexible and generalizable than the standard Neural-ODE. Furthermore, the Neural-GODE enjoys the computational and memory benefits while performing similarly in prediction accuracy compared to ResNets. This is a collaboration work with Drs. Duo Yu and Hongyu Miao.

Nonparametric Bayesian Q-learning for adjusting partial compliance in multi-stage randomized trials

♦ *Indrabati Bhattacharya, Brent Johnson and Ashkan Ertefaie*

University of Rochester

Indrabati.Bhattacharya@URMC.Rochester.edu

Q-learning is a well-known reinforcement learning approach for estimation of optimal dynamic treatment regimes. Existing methods for estimation of dynamic treatment regimes are limited to intention-to-treat analyses—which estimate the effect of randomization to a particular treatment regime without considering the compliance behavior of patients. In this article, we propose a novel Bayesian nonparametric Q-learning approach based on stochastic decision rules for adjusting partial compliance. We consider the popular potential compliance framework, where some potential compliances are latent and need to be imputed. For each stage, we fit a locally weighted Dirichlet process mixture model for the conditional distribution of potential outcomes given the compliance values and baseline covariates. The key challenge is learning the joint distribution of the potential compliances, which we do using a Dirichlet process mixture model. Our approach provides two sets of decision rules: (1) conditional decision rules given the potential compliance values; and (2) marginal decision rules where the potential compliances are marginalized. Extensive simulation studies show the effectiveness of our method compared to intention-to-treat analyses. We apply our method on the Adaptive Treatment For Alcohol and Cocaine Dependence Study (ENGAGE), where the goal is to construct optimal treatment regimes to engage patients in therapy.

Dynamic Topological Data Analysis for Brain Networks

Moo Chung

University of Wisconsin-Madison

mkchung@wisc.edu

We present the novel Wasserstein graph clustering for dynamically changing graphs. The Wasserstein clustering penalizes the topological discrepancy between graphs. The Wasserstein clustering is shown to outperform the widely used k-means clustering. The method is applied in more accurate determination of the state spaces of dynamically changing functional brain networks. The talk is based on arXiv:2201.00087.

Session 3G : Geometric Statistics In Medical Image Computing

Statistical Analysis of Shape Networks

♦ *Anuj Srivastava, Xiaoyang Guo, Aditi Basu Bal and Tom Needham*

Florida State University

anuj@stat.fsu.edu

Imaging data from many applications leads to geometrical structures resembling complex pathways or curvilinear networks. We will call them "shape networks." A prominent example of a shape network is the Brain Arterial Network or BAN in the human brain, which is a complex arrangement of individual arteries, branching patterns, and inter-connectivities. Another example is a road network. Shapes or structures of these objects play an essential role

in characterizing and understanding the functionality of larger systems. One would like tools for statistically analyzing shape networks, i.e., quantifying shape differences, summarizing shapes, comparing populations, and studying the effects of covariates on these shapes. This paper represents and statistically analyzes shape networks as "elastic shape graphs". Each elastic shape graph consists of nodes, or points in 3D, connected by some 3D curves, or edges, with arbitrary shapes. We develop a mathematical representation, a Riemannian metric, and other geometrical tools, such as computations of geodesics, means, covariances, and PCA, for helping analyze elastic shape graphs. We apply this framework to analyzing shapes of BANs taken from 92 subjects. Specifically, we generate shape summaries of BANs, perform shape PCA, and study the effects of age and gender on their shapes. We conclude that age has a clear, quantifiable effect on BAN shapes. Specifically, we find an increased variance in BAN shapes as age increases.

Feature Gradient Flow for Interpretation of Deep Learning Models

P. Thomas Fletcher

University of Virginia

ptf8v@virginia.edu

Recently, deep learning models have shown great promise in medical imaging tasks, such as in making diagnostic predictions. However, such models not made it into standard clinical care, primarily due to a lack of understanding of why a model works and why it fails. In this talk, I will present a recently proposed method for enhancing the interpretability of classifiers, called feature gradient flow. The gradient flow of a model locally defines nonlinear coordinates in the input data space representing the information the model is using to make its decisions. The feature gradient flow then compares a model's gradient flow to the gradients of derived features from the images that are interpretable to humans. We then develop a technique for training neural networks to be more interpretable by adding a regularization term to the loss function that encourages the model gradients to align with those of chosen interpretable features. I will show examples of model interpretability using feature gradient flow on synthetic data as well as data from medical imaging studies.

Nested Homogeneous Spaces: Construction, Learning and Applications

Baba Vemuri

University of Florida

vemuri@ufl.edu

The homogeneous space of a Lie Group G , is a manifold M on which the group G acts transitively. Intuitively, every point in a homogeneous space looks locally alike in the sense of an isometry, diffeomorphism or a homeomorphism. Such spaces are abundant in practice e.g., the vector space of reals, n -sphere, Grassmanian, hyperbolic space, manifold of symmetric positive definite matrices etc. In statistics and machine learning, principal component analysis is the de facto choice for dimensionality reduction and produces nested linear subspaces. In this talk, I will present a recipe for generalizing this concept of producing nested subspaces to homogeneous spaces in general, and show how this general recipe can be embedded into a learning framework. Specific examples of dimensionality reduction and pattern classification using the nested homogeneous space model will be presented for the Grassmanian and the hyperbolic space. In the latter case, the nested hyperbolic space model will be used to develop a nested hyperbolic graph neural network. Experimental results on a variety of synthetic and real

data sets depicting the performance of the models in comparison to the state-of-the-art will be interspersed throughout the presentation.

Integrated Construction of Multimodal Atlases with Structural Connectomes in the Space of Riemannian Metrics

Sarang Joshi

University of Utah
sjoshi@sci.utah.edu

The structural network of the brain, or structural connectome, can be represented by fiber bundles generated by a variety of tractography methods. While such methods give qualitative insights into brain structure, there is controversy over whether they can provide quantitative information, especially at the population level. In order to enable population-level statistical analysis of the structural connectome, we propose representing a connectome as a Riemannian metric, which is a point on an infinite-dimensional manifold. We equip this manifold with the Ebin metric, a natural metric structure for this space, to get a Riemannian manifold along with its associated geometric properties. We then use this Riemannian framework to apply object-oriented statistical analysis to define an atlas as the Fréchet mean of a population of Riemannian metrics. This formulation ties into the existing framework for diffeomorphic construction of image atlases, allowing us to construct a multimodal atlas by simultaneously integrating complementary white matter structure details from DWMRI and cortical details from T1-weighted MRI. We illustrate our framework with 2D data examples of connectome registration and atlas formation. Finally, we build an example 3D multimodal atlas using T1 images and connectomes derived from diffusion tensors estimated from a subset of subjects from the Human Connectome Project.

Session 3H : The Jiann-Ping Hsu Invited Session on Biostatistical and Regulatory Sciences

Covariate-Balancing-Aware Interpretable Deep Learning Models for Treatment Effect Estimation

♦*Kan Chen, Qishuo Yin and Qi Long*

University of Pennsylvania
kanchen@sas.upenn.edu

Estimating treatment effects is of great importance for many biomedical applications with observational data. Particularly, interpretability of the treatment effects is preferable for many biomedical researchers. In this paper, we first give a theoretical analysis and propose an upper bound for the bias of average treatment effect estimation under the strong ignorability assumption. The proposed upper bound consists of two parts: training error for factual outcomes, and the distance between treated and control distributions. We use the Weighted Energy Distance (WED) to measure the distance between two distributions. Motivated by the theoretical analysis, we implement this upper bound as an objective function is minimized by leveraging a novel additive neural network architecture, which combines the expressivity of deep neural network, the interpretability of generalized additive model, the sufficiency of the balancing score for estimation adjustment, and covariate balancing properties of treated and control distributions, for estimating average treatment effects from observational data. Furthermore, we impose a so-called weighted regularization procedure based on non-parametric theory, to obtain some desirable asymptotic properties. The proposed method is illustrated by re-examining the benchmark datasets for causal inference, and it outperforms the state-of-art.

How to Implement the “One Patient, One Vote” Principle under the Framework of Estimand?

♦*Naitee Ting*

Boehringer Ingelheim
naitee.ting@boehringer-ingelheim.com

The scientific foundation of a modern clinical trial is randomization – each patient is randomized to a treatment group, and statistical comparisons are made between treatment groups. Because the study units are individual patients, this “one patient, one vote” principle needs to be followed – both in study design, and in data analysis. From the physicians’ point of view, each patient is equally important, they need to be treated equally in data analysis. It is critical that statistical analysis should respect design, and study design is based on randomization. Hence from both statistical and medical points of view, data analysis needs to follow this “one patient, one vote” principle. Under ICH E9 (R1), five strategies are recommended to establish “estimand”. This manuscript discusses how to implement these strategies using the one patient, one vote principle.

Cox Model for Weibull Survival Data

Mario Keko, ♦Marwan Alsharman, Djhenne Dalmacy, Lili Yu

Georgia Southern University
ma15387@georgiasouthern.edu

Survival analysis has a great application in the analysis of time to event data. Among different approaches, semiparametric Cox Hazards Model has advantages in terms of implementation, flexibility, and interpretation, given its assumptions are met. In this study, we will investigate the performance of the Cox model, compared with the parametric model, for Weibull survival data. We will compare the coefficient estimation and baseline hazards estimation from the parametric Weibull model with those from the Cox model. The simulations show the Cox model performs well for the coefficient estimation, but not well for baseline hazards estimation.

An Application of the Cure Model to A Cardiovascular Clinical Trial

♦*Varadan Sevilimedu, S Ma, P Hartigan, TC Kyriakides*

Memorial Sloan Kettering Cancer Center
varadan.sevilimedu@gmail.com

Intermediate events play an important role in determining the risk of acquiring disease over time, thus making them an important entity to be studied in survival analysis. Myocardial infarction (MI) is one such disease whose hazards are also dependent upon the occurrence of an intermediate event such as acute coronary syndrome (ACS). The study of the role that ACS plays in altering the hazards of MI becomes complicated when there is a cure fraction in the population. Data from the Clinical Outcomes Utilizing Revascularization and Aggressive Drug Evaluation (COURAGE) trial provide the scenario where the existence of a cure fraction is highly likely. In this study we assess the role of ACS in altering the pathway towards developing an MI, in the presence of a cure fraction. We adapt a non-parametric maximum likelihood estimation approach to estimate the regression coefficients of this multi-part cure model. Simulation studies show that the estimates have good asymptotic properties and are robust to variation in data. In addition, we also utilize the dataset to explore the use of a proportionality constraint to aid in reduction of dimensionality. The analysis yielded some particularly novel findings that could prove useful in guiding clinical practice.

Session 4A : Recent Developments For Causal Inference: Theory, Method, And Application

Calibrated Optimal Decision Making with Multiple Data Sources and Limited Outcome

♦ *Hengrui Cai, Wenbin Lu and Rui Song*

North Carolina State University
hcai5@ncsu.edu

We consider the optimal decision-making problem in a primary sample of interest with multiple auxiliary sources available. The outcome of interest is limited in the sense that it is only observed in the primary sample. In reality, such multiple data sources may belong to heterogeneous studies and thus cannot be combined directly. This paper proposes a new framework to handle heterogeneous samples and address the limited outcome simultaneously through a novel calibrated optimal decision-making method, by leveraging the common intermediate outcomes in multiple data sources. Specifically, our method allows the baseline covariates across different samples to have either homogeneous or heterogeneous distributions. Under the equal conditional means of intermediate outcomes in different samples given baseline covariates and the treatment information, we show that the proposed estimator of the conditional mean outcome is asymptotically normal and more efficient than using the primary sample solely. Extensive experiments on simulated datasets demonstrate empirical validity and improved efficiency using our approach, followed by a real application to electronic health records.

A Focusing Framework for Testing Bi-Directional Causal Effects with GWAS Summary Data

Ting Ye

University of Washington
tingyel@uw.edu

Mendelian randomization (MR) is a powerful method that uses genetic variants as instrumental variables (IVs) to infer the causal effect of a modifiable exposure on an outcome. Although recent years have seen many extensions of basic MR methods to be robust to certain violations of assumptions, few methods were proposed to infer bi-directional causal relationships, especially for phenotypes with limited biological understandings. The presence of horizontal pleiotropy adds another layer of complexity. In this article, we show that assumptions for common MR methods are often impossible or too stringent in the existence of bi-directional relationships. We then propose a new focusing framework for testing bi-directional causal effects between two traits with possibly pleiotropic genetic variants. Our proposal can be coupled with many state-of-art MR methods. We provide theoretical guarantees on the Type I error and power of the proposed methods. We demonstrate the robustness of the proposed methods using several simulated and real datasets.

Sensitivity Analysis of Individual Treatment Effects: A Robust Conformal Inference Approach

Ying Jin¹, ♦ Zhimei Ren² and Emmanuel Candès¹

¹Stanford University

²University of Chicago
zmren@uchicago.edu

We propose a model-free framework for sensitivity analysis of individual treatment effects (ITEs), building upon ideas from conformal inference. For any unit, our procedure reports the Γ -value, a number which quantifies the minimum strength of confounding needed to explain away the evidence for ITE. Our approach rests on the reliable predictive inference of counterfactuals and ITEs in situations

where the training data is confounded. Under the marginal sensitivity model of Tan (2006), we characterize the shift between the distribution of the observations and that of the counterfactuals. We first develop a general method for predictive inference of test samples from a shifted distribution; we then leverage this to construct covariate-dependent prediction sets for counterfactuals. No matter the value of the shift, these prediction sets (resp. approximately) achieve marginal coverage if the propensity score is known exactly (resp. estimated). We describe a distinct procedure also attaining coverage, however, conditional on the training data. In the latter case, we prove a sharpness result showing that for certain classes of prediction problems, the prediction intervals cannot possibly be tightened. We verify the validity and performance of the new methods via simulation studies and apply them to analyze real datasets.

Causal inference of time-varying effects in non-stationary time series using mobile health data

♦ *Xiaoxuan Cai¹, Jukka-Pekka Onnela², Justin Baker³, Habib Rahimi-Eichi³ and Linda Valeri¹*

¹Columbia University

²Harvard University

³McLean Hospital
nmcaixiaoxuan@gmail.com

Mobile technology (e.g., mobile phones and wearable devices) enables unprecedented continuous monitoring of an individual's behavior, social interactions, symptoms, and other health conditions. Continuous measurements of personal data results in the emergence of multivariate time series of outcomes, exposures, and covariates in a N-or-1 study. Additionally, it enables estimations of a broader class of causal quantities that can be used to describe how time-varying exposure influences outcome in the short and long term, as well as how this influence varies over time. Popular methods for univariate time series or longitudinal data emphasize the contemporaneous effect or the effect of a complete treatment history, and they presume stationary time series or time-invariant treatment effects. However, these methods are inadequate to capture the full picture of exposure/treatment effects in the short and long term of mobile health data, and are not designed to capture the time-varying behavior of causal quantities in non-stationary multivariate time series. We propose a set of causal estimands for (potentially non-stationary) multivariate time series in N-of-1 studies in order to systematically summarize how time-varying exposures affect outcomes in the short and long term, and demonstrate their identification via the g-formula in the presence of exposure- and outcome-covariate feedbacks. The g-formula employs an innovative state space model to capture the time-varying behavior of treatment effects. We demonstrate our proposed estimands and the identification of these time-varying effects using a smartphone observational study of bipolar patients, in which we examine both the short- and long-term effects of social support on patients' mood and how these effects vary over time, while accounting for feedback between exposure, outcome, and covariates.

Session 4B : High-Dimensional Statistical Inference For Big Complicated Data

Anti-Concentration of Suprema of Gaussian Processes with Applications to High-Dimensional CLTs

Alexander Giessing
University of Washington
giessing@uw.edu

We present a two-sided anti-concentration inequality for the Levy concentration function of suprema of separable, sample bounded Gaussian processes. Lower and upper bounds on the Levy concentration function differ only by a multiplicative constant and depend on the variance of the supremum. Such anti-concentration inequalities play an important role in establishing bounds on the Kolmogorov distance between suprema of empirical and Gaussian processes when the function classes are not Donsker. As an application, we derive dimension-free Berry-Esseen-type bounds for high-dimensional CLTs and the conditional multiplier bootstrap.

Multiple-Splitting Projection Test for High-Dimensional Mean Vectors

Wanjuan Liu¹, ♦Xiufan Yu² and Runze Li³

¹LinkedIn Corporation

²University of Notre Dame

³Penn State University
xiufan.yu@nd.edu

We propose a multiple-splitting projection test (MPT) for one-sample mean vectors in high-dimensional settings. The idea of projection test is to project high-dimensional samples to a 1-dimensional space using an optimal projection direction such that traditional tests can be carried out with projected samples. However, estimation of the optimal projection direction has not been systematically studied in literature. In this work, we bridge the gap by proposing a consistent estimation via regularized quadratic optimization. To retain type I error rate, we adopt a data-splitting strategy when constructing test statistics. To mitigate the power loss due to data-splitting, we further propose a test via multiple splits to enhance the testing power. We show that the p-values resulted from multiple splits are exchangeable. Unlike existing methods which tend to conservatively combine dependent p-values, we develop an exact level α test that explicitly utilizes the exchangeability structure to achieve better power. Numerical studies show that the proposed test well retains the type I error rate and is more powerful than state-of-the-art tests.

Two-sample hypothesis testing of multiple-network data

♦Yinqiu He¹, Xuming He², Ji Zhu² and Gongjun Xu²

¹Columbia University

²University of Michigan
yinqiu.he@columbia.edu

Multiple-network data has attracted increasing attention recently, where the data are recorded as symmetric matrices, and each matrix encodes an individual network structure. Such data arises frequently in various scientific fields such as the analyses of brain connectivity and gene interactions. In these studies, it is of great interest to compare the means of two populations of networks. In this work, we propose a hypothesis testing procedure when we are interested in a given area of networks that may have multiple signals functioning together. We establish asymptotic results for the proposed test under general moment conditions and validate the test under a variety of popular network models. We further demonstrate the efficacy of the proposed test under simulation studies and the analysis of a brain dataset.

Doubly Debiased Lasso: High-Dimensional Inference under Hidden Confounding

♦Zijian Guo¹, Domagoj Cevic² and Peter Buhlmann²

¹Rutgers

²ETH, Zurich
zijguo@stat.rutgers.edu

Inferring causal relationships or related associations from observational data can be invalidated by the existence of hidden confounding. We focus on a high-dimensional linear regression setting, where the measured covariates are affected by hidden confounding and propose the *Doubly Debiased Lasso* estimator for individual components of the regression coefficient vector. Our advocated method simultaneously corrects both the bias due to estimation of high-dimensional parameters as well as the bias caused by the hidden confounding. We establish its asymptotic normality and also prove that it is efficient in the Gauss-Markov sense. The validity of our methodology relies on a dense confounding assumption, i.e. that every confounding variable affects many covariates. The finite sample performance is illustrated with an extensive simulation study and a genomic application.

Session 4C : New Fronts In Joint Modeling And Machine Learning

Joint modeling for longitudinal and interval censored survival data

Ding-Geng Chen

Arizona State University
Ding-Geng.Chen@asu.edu

Joint models for longitudinal and survival data are a class of models that jointly analyze an outcome repeatedly observed over time (such as a bio-marker) and associated event times. The joint modelling framework has mainly focused on right-censored time-to-event data in the survival outcome. In this talk, we discuss this joint modeling framework for interval-censored survival data with a real example from a cardiology multicenter clinical trial (IMPI trial).

Heterogeneous Data Integration And The Predictive Ability of Cancer Survival Models

Yi Guo

Health Outcomes & Biomedical Informatics, University of Florida
yigu@ufl.edu

To improve cancer survival, a deep understanding of the associated multi-level contributory factors is much needed. However, prior research on cancer survival has primarily focused on factors from the individual level due to the limited availability of integrated datasets. In this study, we sought to examine how data integration impacts the performance of cancer survival models. We linked data from four different sources and evaluated the performance of Cox proportional hazard models for breast, lung, and colorectal cancers under three common data integration scenarios. We showed that adding additional contextual-level predictors to survival models through linking multiple datasets improved model fit and performance. We also showed that different representations of the same variable or concept have differential impacts on model performance.

Regression Analysis of Mixed Panel-Count Data with Application to Cancer Studies

♦Yimei Li¹, Liang Zhu², Lei Liu³ and Leslie Robison⁴

¹St Jude Children's Research Hospital

²Eisai

³Washington University

⁴St. Jude Children's Research Hospital
Yimei.Li@STJUDE.ORG

Both panel-count data and panel-binary data are common data types in recurrent event studies. Because of inconsistent questionnaires or missing data during the follow-ups, mixed data types need to be addressed frequently. A recently proposed semiparametric approach

uses a proportional means model to facilitate regression analyses of mixed panel-count and panel-binary data. This method can use all available information regardless of the record type and provide unbiased estimates. However, the large number of nuisance parameters in the nonparametric baseline hazard function makes the estimating procedure very complicated and time-consuming. We approximated the baseline hazard function to simplify the estimating procedure. Simulation studies showed that our method performed similarly to that of the previous semiparametric likelihood-based method, but with much faster speed. Approximating the baseline hazard not only reduced the computational burden but also made it possible to implement the estimating procedure in a standard software, such as SAS.

Joint modeling in presence of informative censoring in palliative care studies

♦*Quran Wu¹, Michael Daniels², Areej Jawahri³, Marie Bakitas⁴ and Zhigang Li¹*

¹Department of Biostatistics, University of Florida

²Department of Statistics, University of Florida

³Department of Oncology, Massachusetts General Hospital

⁴School of Nursing, University of Alabama at Birmingham
wuc@nhlbi.nih.gov

Joint modeling of longitudinal data such as quality of life data and survival data is important for palliative care researchers to draw efficient inference because it can account for the associations between those two types of data. Modeling quality of life on a retrospective time scale from death time makes it convenient for investigators to interpret the analysis results of palliative care studies with relatively short life expectancy. However, censoring of death times, especially informative censoring such as informative dropouts, poses challenges for modeling quality of life on a retrospective time scale. We develop a novel joint modeling approach that can address the challenge by allowing informative censoring events to be dependent on patients' quality of life through a random effect. There are three submodels in our approach: a linear mixed effect model for the longitudinal quality of life, a frailty model for the death time and another frailty model for the informative censoring time. In addition to improving the precision of estimates, our approach can provide unbiased estimates for making valid inference by appropriately modeling the informative censoring time. Model performance is assessed with a simulation study in comparison with existing approaches. A real-world study is presented to showcase the application of the new approach.

Session 4D : Knowledge-Guided Machine Learning And Statistical Modeling In Longitudinal Studies With Survival Endpoints

Design and Analysis of a Multi-Platform Trial of Patients Hospitalized for COVID-19

♦*Eric Leifer¹, Lucy Kornblith, Jeffrey Berger, Lana Castellucci, Michael Farkouh, Ewan Goligher, Patrick Lawler and Scott Berry*

¹NIH/NHLI

leifere@nhlbi.nih.gov

Patients hospitalized for COVID-19 are predisposed to development of serious thromboembolic complications and associated poor outcomes including organ failure and death. In response to this, three clinical trial platforms were combined to conduct an international, open-label, adaptive, clinical trial which randomized patients hospitalized for COVID-19 to receive either therapeutic or

usual care pharmacologic heparin anticoagulation for thromboprophylaxis. The primary endpoint was the number of organ-support free days survival through 21 days. Several design choices were made. Patients were first stratified according to whether they required intensive level of care at baseline, so called severe patients. Patients who did not require intensive level of care at baseline, so called moderate patients, were further stratified by their baseline D-dimer level. There was the potential for separate trial conclusions in severe and moderate patients according to D-dimer level. To facilitate rapid reporting of results, a Bayesian cumulative logistic model, also known as a proportional odds model, was used. This model included hierarchical parameters to allow for dynamic borrowing of outcome data between severe and moderate patients, with greater borrowing when the outcome data was more similar. The proportional odds ratio treatment effect parameter was separately monitored for the severe and moderate patients. Monitoring was done using Bayesian stopping guidelines for efficacy and futility. The flexibility in the trial design allowed for rapid reporting of results, which did differ between severe and moderate patients. The analysis of those results will be presented.

Knowledge-Guided Model Building and Estimation with Time-to-Event Outcomes and Longitudinal Covariates

♦*Colin O. Wu¹, Xiaoyang Ma and Xin Tian*

¹Division of Intramural Research

wuc@nhlbi.nih.gov

In large epidemiological studies, a comprehensive analysis of a disease process often involves several statistical sub-models that describe different aspects of the covariates and the disease outcome. A final prediction model for the disease can be constructed by incorporating all the influential covariates and sub-models, so that meaningful clinical interpretations could be obtained. Existing statistical machine learning methods lack a systematic approach for incorporating all these influential covariates and sub-models in a biologically meaningful way. We describe a knowledge-guided machine learning procedure to construct a comprehensive statistical model for predicting the distributions of time-to-event outcomes with longitudinal covariates. This procedure combines several statistical machine learning approaches with the biomedical knowledge established in the literature. We apply our procedure to the Coronary Artery Risk Development in Young Adults (CARDIA) study and demonstrate that this procedure leads to novel insights into the effects of longitudinal risk factors on the distributions of incident cardiovascular disease (CVD). We demonstrate the appropriateness of our procedure through a simulation study.

Dynamic Risk Prediction Triggered by Intermediate Events Using Survival Tree Ensembles

♦*Yifei Sun¹, Sy Han Chiou², Colin Wu³, Meghan McGarry⁴ and Chiung-Yu Huang⁴*

¹Columbia University

²University of Texas at Dallas

³National Heart, Lung, and Blood Institute

⁴University of California San Francisco

ys3072@cumc.columbia.edu

With the availability of massive amounts of data from electronic health records and registry databases, incorporating time-varying patient information to improve risk prediction has attracted great attention. To exploit the growing amount of predictor information over time, we develop a unified framework for landmark prediction using survival tree ensembles, where an updated prediction can be performed when new information becomes available. Compared to

conventional landmark prediction with fixed landmark times, our methods allow the landmark times to be subject-specific and triggered by an intermediate clinical event. Moreover, the nonparametric approach circumvents the thorny issue of model incompatibility at different landmark times. In our framework, both the longitudinal predictors and the event time outcome are subject to right censoring, and thus existing tree-based approaches cannot be directly applied. To tackle the analytical challenges, we propose a risk-set-based ensemble procedure by averaging martingale estimating equations from individual trees. Extensive simulation studies are conducted to evaluate the performance of our methods. The methods are applied to the Cystic Fibrosis Patient Registry (CFPR) data to perform dynamic prediction of lung disease in cystic fibrosis patients and to identify important prognosis factors.

Dealing With Competing Risks in Clinical Trials

James Troendle

NIH

troendlj@mail.nih.gov

We investigate different primary efficacy analysis approaches for a 2-armed randomized clinical trial when interest is focused on a time to event primary outcome that is subject to a competing risk. We extend the work of Freidlin and Korn (2005) by considering estimation as well as testing and by simulating the primary and competing events' times from both a cause-specific hazards model as well as a joint subdistribution-cause-specific hazards model. We show that the cumulative incidence function can provide prognostic information for a particular patient but is not advisable for the primary efficacy analysis. Instead, it is preferable to fit a Cox model for the primary event which treats the competing event as an independent censoring. This is reasonably robust for controlling type I error and treatment effect bias with respect to the true primary and competing events' cause-specific hazards model, even when there is a shared, moderately prognostic, unobserved baseline frailty for the primary and competing events in that model. However, when it is plausible that a strongly prognostic frailty exists, combining the primary and competing events into a composite event should be considered. Finally, when there is an a priori interest in having both the primary and competing events in the primary analysis, we compare a bivariate approach for establishing overall treatment efficacy to the composite event approach. The ideas are illustrated by analyzing the Women's Health Initiative clinical trials sponsored by the National Heart, Lung, and Blood Institute.

Session 4E : Robust Information Integration From Multiple Studies In Clinical And Biomedical Research

On multi-site collaboration, data sharing, and analytic strategy in medical research

♦Jing Huang¹, Rui Duan² and Yong Chen¹

¹University of Pennsylvania

²Harvard University

jing14@penncmedicine.upenn.edu

Collaboration is a vital component of medical research. Over many decades, great efforts have been made by cross-discipline stakeholders to promote national and international collaborations in both clinical trials and observational studies. Research networks, consortia, and collaborative programs were founded with growing numbers of participating organizations under a shared vision to accelerate discovery and improve health care by integrating data, protocols, and resources. A benefit of such a multi-site collaboration is the gain of

a larger and diverse coverage of the study population, which could lead to a better chance of knowledge discovery. Data sharing is a critical component of multi-site collaborations. Depending on the contexts, however, the method of data sharing across sites varies in different ways. In this study, we conducted simulation studies to empirically quantify the difference in study bias and logistical burden of three strategies: mega-analysis, meta-analysis, and distributed data analysis. We mimicked a common setup in medical research where a cross-sectional study was conducted to evaluate the association between a binary exposure and a binary outcome.

Integrating summary information from many external studies with heterogeneous populations

Peisong Han

University of Michigan

peisong@umich.edu

For an internal study of interest, information provided by relevant external studies can be useful to improve the efficiency for parameter estimation in model building, and the external information is oftentimes in summary form. When information is available from possibly many external studies, extra care is needed due to inevitable study population heterogeneity. The information from studies with populations different from the internal study may harm model fitting by introducing estimation bias. We allow the number of external studies that can be considered for possible information integration to increase with the internal sample size, and develop an effective method that integrates only the helpful information for efficiency improvement without introducing bias. Using this method, we study the change of mental health for individuals with bipolar disorder during COVID-19 pandemic, by integrating summary information from relevant existing large-scale studies.

Data Integration Methods Targeting Underrepresented Populations in Precision Medicine

Rui Duan

Harvard University

rduan@hsph.harvard.edu

The limited representation of minorities and disadvantaged populations in large-scale clinical and genomics research has become a barrier to translating precision medicine research into practice, which may further exacerbate known health disparities, and there is an urgent need for statistical and machine learning methods to address biases and unfairness caused by such lack of representation. To address this problem, we propose two data integration strategies: (1) leveraging the shared knowledge from diverse populations, and (2) integrating larger bodies of data from multiple healthcare institutions. We develop a transfer learning framework to transfer shared knowledge learned across multiple populations to an underrepresented population so that comparable performance can be obtained with much fewer data. We also develop federated learning algorithms to enable the integration of data across multiple healthcare organizations without directly sharing individual-level. We show that our methods improve the estimation and prediction accuracy in underrepresented populations, and can reduce the gap in model performance across populations. We provide theoretical analyses revealing how the estimation accuracy of the proposed method is influenced by communication budgets, privacy restrictions, and heterogeneity across populations. We demonstrate the feasibility and validity of our methods through numerical experiments and apply them to construct prediction models for multiple complex diseases and traits using data from the electronic Medical Records and Genomics (eMERGE) Network.

Integrated Analysis of Randomized Clinical Trials with Real-World Data

♦Xiaofei Wang¹, Dasom Lee² and Shu Yang²

¹Duke University

²NC State University

xiaofei.wang@duke.edu

In this talk, we exploit the complementing features of randomized clinical trials (RCT) and real world evidence (RWE) to estimate the average treatment effect of the target population. We will review existing methods in conducting integrated analysis of binary, continuous and survival data from RCTs and RWEs. We will then discuss in detail new calibration weighting estimators that are able to calibrate the covariate information between RCTs and RWEs. We will briefly review asymptotic results under mild regularity conditions, and confirm the finite sample performances of the proposed estimators by simulation experiments. In a comparison of existing methods, we illustrate our proposed methods to estimate the effect of adjuvant chemotherapy in early-stage resected non-small-cell lung cancer integrating data from a RCT and the National Cancer Database.

Session 4F : Statistical Innovation In Complex And High Dimensional Data

Change detection in certain random intensity-driven point processes through repeated testing

Moinak Bhaduri

Bentley University

mbhaduri@bentley.edu

As we surface, probably momentarily, from the pandemic, other crises thwart normalcy: appalling inequality, climate calamity, distressed refugees, upped possibilities of a fresh Cold War. The enduring motif of our time is incessant chaos. Frequently, that chaos results when one type of a stationary system gives way to another. Change detection is mainly about estimating these points of deviation. In case a Poisson-type point process carries the system forward, I'll offer a brand of detection algorithms, engineered through permutations of trend switched statistics and a judicious application of false discovery rate control. Certain members of this family that remain asymptotically consistent and close to the ground truth (evidenced through some Hausdorff-similarity) are isolated to pinpoint estimated change locations. Efficient forecasting proves to be a natural corollary. Change point-based clustering tools will also be examined. I'll describe how such analyses offer concrete definitions to vague objects like Covid "waves" and measure their enormity. Suggested quick reading: *Contrary Currents*, CHANCE, 35:1, 26-33, DOI: 10.1080/09332480.2022.2039025

A nonparametric multi-sample test for high-dimensional compositional data with applications to the human microbiome

Qingyang Zhang

University of Arkansas

qz008@uark.edu

Compositional data refer to the data that lie on a simplex, which are common in many scientific domains such as genomics, geology and economics. In this work, we consider a general problem of testing for the compositional difference between K populations. Motivated by microbiome and metagenomics studies, where the data are often over-dispersed and high-dimensional, we formulate a well-posed hypothesis from a Bayesian point of view and suggest a non-parametric test based on inter-point distance to evaluate statistical significance. Unlike most existing tests for compositional data, our

method does not rely on any data transformation, sparsity assumption or regularity conditions on the covariance matrix, but directly analyzes the compositions. Our simulation studies and real data applications demonstrate that the proposed test is more sensitive to the compositional difference than the mean-based method, especially when the data are over-dispersed or zero-inflated. The proposed test is easy to implement and computationally efficient, facilitating its application to large-scale datasets.

Minimum discrepancy approach for dimension reduction by filtered feature

Pei Wang

Miami University

wangp33@miamioh.edu

The minimum discrepancy approach is useful in sufficient dimension reduction (SDR). In this article, we develop a novel SDR method through a minimum discrepancy approach using filtered features. To obtain the sparse solution in the ultra-high dimensional data, a regularization method is proposed. The asymptotic results are established and the estimation method for determining structural dimension is provided. We demonstrate the efficacy of our method through extensive simulations and a real data example.

Nonparametric Mixture Model: Application in Contaminated Trials

Zi Ye

Lehigh University

ziy421@lehigh.edu

In personalized medicine, investigating the differential effect of treatments in groups defined by patient characteristics is of paramount importance. In randomized clinical trials setting, participants are first classified by using diagnostic tools, but such classifiers may not be perfectly accurate. The issue of diagnostic misclassification has recently become prominent and has been shown to produce severely biased estimations of treatment effect. In this talk, we analyze this problem in a pre-post design. For ordinal, discrete or skewed outcomes, we develop a fully nonparametric method for estimating and testing treatment effect. Consistent estimators and asymptotic distributions are provided for the misclassification error rates as well as the treatment effect. Simulations study is conducted to compare the new method with traditional methods. The results show significant advantages of the proposed methods in terms of bias reduction, coverage probability and power.

Session 4G : Enhance Decision Making In Early Oncology Studies To Expedite Drug Development

A Bayesian hierarchical monitoring design for phase II cancer clinical trials: Incorporating information on response duration

♦Jian Wang¹, Jing Ning¹, Junsheng Ma¹, Chunyan Cai² and Naval Daver¹

¹The University of Texas MD Anderson Cancer Center

²Marketplace Data Science, Uber

jianwang@mdanderson.org

To screen a new treatment by evaluating its preliminary therapeutic effect, futility monitoring rules are commonly used in phase II cancer clinical trials to make "go/no-go" decisions timely and efficiently. These futility monitoring rules are usually focused on a single outcome (eg, response rate), although a single outcome may not adequately determine the efficacy of the experimental treatment. To address this issue, we propose Bayesian hierarchical futility monitoring rules to consider both the response rate and duration. The

first level of monitoring evaluates whether the response rate provides evidence that the experimental treatment is worthy of further evaluation. If the evidence from the response rate does not support continuing the trial, the second level monitoring rule, based on the duration of response, will be triggered. If both stopping rules are satisfied, the trial will be stopped for futility. We conducted simulation studies to evaluate the operating characteristics of the proposed monitoring rules and compared them to those of standard method. We illustrated the proposed design with a single arm phase II cancer clinical trial to assess the safety and efficacy of a combined treatment in patients with relapsed/refractory acute myeloid leukemia.

Bayesian Interim Monitoring for Faster Decision-Making in Early Phase Trials

Victoria Chang, Kathy Zhang and ♦Gaohong Dong

BeiGene

victoria.chang@beigene.com

The efficacy interim monitoring with either predictive probability or predictive probability of success (PPOS) are widely used in Phase I or Phase II trials. Both are Bayesian approaches which are efficient/flexible and can be easily applied to early phase oncology trial with binary efficacy endpoint (i.e. ORR). In this presentation, the comparison between two approaches will be demonstrated and some interesting results of operation characteristics will be illustrated through simulations.

Session 4H : Design And Analysis Of Computer Experiments

Modeling and Active Learning for Experiments with Quantitative-Sequence Factors

Abhyuday Mandal

University of Georgia

amandal@stat.uga.edu

A new type of experiment which targets on finding optimal quantities of a sequence of factors is drawing much attention in medical science, bio-engineering and many other disciplines. Such studies require simultaneous optimization for both quantities and sequence-orders of several components, which is defined as a new type of factors: quantitative-sequence (QS) factors. Due to the large and semi-discrete solution spaces in such experiments, it is non-trivial to efficiently identify the optimal (or near optimal) solutions using only a few experimental trials. To address this challenge, we propose a novel active learning approach, named as QS-learning, to enable effective modeling and efficient optimization for experiments with QS factors. The QS-learning consists of three parts: a novel mapping-based additive Gaussian process (MaGP) model, an efficient global optimization scheme (QS-EGO), and a new class of optimal designs (QS-design) for collecting initial data. Theoretical properties of the proposed method are investigated and techniques on optimization using analytical gradients are developed. The performance of the proposed method is demonstrated via a real drug experiment on lymphoma treatment and several simulation studies.

Lioness Algorithm for Finding Optimal Design of Experiments

♦Hongzhi Wang, Qian Xiao and Abhyuday Mandal

University of Georgia

hongzhi.wang25@uga.edu

In modern science and engineering applications, optimal designs of experiments, which target on maximizing the information gained from the limited data, are widely used. Depending on the physical

constraints and experimental requirements, various types of optimal designs are used, including D-optimal designs, space-filling designs, orthogonal designs and order-of-addition designs. It is very challenging to construct optimal designs with flexible sizes. Current literature often focus on a single type of designs. In this work, we propose a new nature-inspired metaheuristics optimization algorithm, called the Lioness algorithm (LA), which works efficiently on many different types of design problems. Simulation studies show that the LA outperforms several state-of-the-art methods in terms of both statistical performances and computational resources.

A Simulation Optimization Approach for Sequential Accelerated Life Testing via Approximate Bayesian Inference

Ye Chen¹, ♦Qiong Zhang², Mingyang Li³ and Wenjun Cai⁴

¹Virginia Commonwealth University

²Clemson University

³USF

⁴Virginia Tech

qiongz@clemson.edu

Approximate Bayesian inference has been proposed to construct computationally tractable statistical learning procedures for incomplete or censored data. In this talk, I will discuss a sequential model-updating procedure via approximate Bayesian inference for the Log-normal model with censored observations. We show that the proposed procedure leads to a consistent model parameter estimation. The developed model updating procedure also enables a closed-form expression of a sequential design criterion. The proposed procedure is applied to accelerated life testing experiments, which aim at determining the material alternative with the best reliability performance.

Optimal Crossover Designs for Generalized Linear Models

♦Jeevan Jankar¹, Abhyuday Mandal and Jie Yang²

¹University of Georgia

²University of Georgia

hongzhi.wang25@uga.edu

We identify locally D-optimal crossover designs for generalized linear models. We use generalized estimating equations to estimate the model parameters along with their variances. We discuss a real example of multiple-treatment crossover experiments using Latin square designs. We derive a corresponding general equivalence theorem for crossover designs under generalized linear models. Using a simulation study, we show that a two-stage design with our locally D-optimal design at the second stage is more efficient than the uniform design, especially when the responses from the same subject are correlated.

Session 5A : Statistical Methodologies In Causal Inference With Application In Drug Development

Time and Causality: Learning Causal Structures from Longitudinal Data

Siyi Deng¹, ♦Jiarui Lu² and Dong Xi³

¹Cornell University

²Novartis pharmaceuticals corporation

³Gilead Sciences

jiarui.lu@novartis.com

Understanding disease mechanisms is crucial to drug discovery. Traditional exploratory data analyses often focus on association instead of causation (or cause-effect relationships). Causal graphical models become more and more popular because of their ability to identify causal relationships based on association measures.

With longitudinal data, time is often considered to provide a natural cause-effect relationship, i.e., variables observed at the current time point should cause the same variables in the future but should not cause those observed in the past. In this presentation, we introduced a novel methodology for estimating causal relationships based on longitudinal data. This method addresses the within-subject dependency as well as the cross-time causal relationships via a modified PC algorithm with constraints. A Gaussian Copula model is used to fit the data with mixed types (continuous, binary, etc.). We also proposed a condition called subgraph consistency to protect the cause-effect relationships at the current time point from being affected by future variables. We also conduct extensive simulations to show the validity of our method in recovering the true underlying causal structure.

Minimax optimal subgroup identification

♦ *Matteo Bonvini*¹, *Edward H. Kennedy*¹ and *Luke J. Keele*²

¹Carnegie Mellon University

²University of Pennsylvania
mbonvini@andrew.cmu.edu

Quantifying treatment effect heterogeneity is a crucial task in many areas of causal inference, e.g. optimal treatment allocation and estimation of subgroup effects. We study the problem of estimating the level sets of the conditional average treatment effect (CATE), identified under the no-unmeasured-confounders assumption. Given a user-specified threshold, the goal is to estimate the set of all units for whom the treatment effect exceeds that threshold. For example, if the cutoff is zero, the estimand is the set of all units who would benefit from receiving treatment. Assigning treatment just to this set represents the optimal treatment rule that maximises the mean population outcome. Similarly, cutoffs greater than zero represent optimal rules under resource constraints. Larger cutoffs can also be used for anomaly detection, i.e., finding which subjects are most affected by treatments. Being able to accurately estimate CATE level sets is therefore of great practical relevance. The level set estimator that we study follows the plug-in principle and consists of simply thresholding a good estimator of the CATE. While many CATE estimators have been recently proposed and analysed, how their properties relate to those of the corresponding level set estimators remains unclear. Our first goal is thus to fill this gap by deriving the asymptotic properties of level set estimators depending on which estimator of the CATE is used. Next, we identify a minimax optimal estimator in a model where the CATE, the propensity score and the outcome model are Holder-smooth of varying orders. We consider data generating processes that satisfy a margin condition governing the probability of observing units for whom the CATE is close to the threshold. We investigate the performance of the estimators in simulations and illustrate our methods on a dataset from REFLUX, a multi-center study that aimed to compare the effectiveness of surgery to treat Gastro-Oesophageal Reflux Disease.

A Bayesian Machine Learning Approach for Estimating Heterogeneous Survivor Causal Effects: Applications to a Critical Care Trial

♦ *Xinyuan Chen*¹, *Michael O. Harhay*², *Guangyu Tong*³ and *Fan Li*³

¹Mississippi State University

²University of Pennsylvania

³Yale University
xchen@math.msstate.edu

Assessing heterogeneity in the effects of treatments has become increasingly popular in the field of causal inference and carries im-

portant implications for clinical decision-making. While extensive literature exists for studying treatment effect heterogeneity when outcomes are fully observed, there has been limited development of tools for estimating heterogeneous causal effects when patient-centered outcomes are truncated by a terminal event, such as death. Due to mortality occurring during study follow-up, the outcomes of interest are unobservable, undefined, or not fully observed for specific subgroups of participants, therefore requiring the principal stratification framework to draw valid causal conclusions. Motivated by the Acute Respiratory Distress Syndrome Network (ARDSNetwork) ARDS respiratory management (ARMA) trial, we developed a flexible Bayesian machine learning approach to estimate the average causal effect and heterogeneous causal effects among the always-survivors stratum when clinical outcomes are subject to truncation. We adopted Bayesian additive regression trees (BART) to flexibly specify separate models for the potential outcomes and latent strata membership. In the analysis of the ARMA trial, we found that the low tidal volume treatment had an overall benefit for participants sustaining acute lung injuries on the outcome of time to returning home, but substantial heterogeneity in treatment effects among the always-survivors, driven most strongly by sex and the alveolar-arterial oxygen gradient at baseline (a physiologic measure of lung function and source of hypoxemia). These findings illustrate how the proposed methodology could guide the prognostic enrichment of future trials in the field. We also demonstrated through a simulation study that our proposed Bayesian machine learning approach outperforms other parametric methods in reducing the estimation bias in both the average causal effect and heterogeneous causal effects for always-survivors.

Application of the causal inference in estimands for a principal stratum in clinical trials

Yongming Qu

Eli Lilly and Company
QU.YONGMING@LILLY.COM

Presently, randomized clinical trials remain the gold standard to evaluate the efficacy and safety of a new treatment. Randomization provides the basis for drawing statistically valid causal inference for the treatment effect. However, patients may not always take the medication to which they were randomized, thus preventing the observation of their potential outcomes under the assigned treatments. Understanding the treatment effect for those who would adhere to one or both treatment groups is an important clinical question. The traditional naive per-protocol analysis is not corresponding a causal estimand and should not be conducted. In this presentation, we will define the causal estimands for the adherers and derive the corresponding estimators based on the principal stratification framework. The estimates can be obtained from the R package "adace" or using multiple imputation.

Session 5B : Recent Developments Of Dimension Reduction In Integrating Big And Complex Data

Nonlinear envelope model

♦ *Bing Li*¹, *Zhihua Su*² and *Dennis Cook*³

¹Penn State University

²University of Florida

³University of Minnesota
bx19@psu.edu

We propose a nonlinear envelope model where the regression function is a member of a reproducing kernel Hilbert space. This ap-

proach builds a parsimonious structure on the regression function, the covariance operator of the predictor, and the response covariance matrix, thus achieving substantially enhanced estimation and prediction. We impose a universal kernel on the predictor and a linear kernel on the response, which fully capture the statistical dependence in the nonparametric regression model. The predictor envelope is defined through the invariance lattice of the covariance operator of X , and the response envelope through the invariance lattice of covariance matrix of Y . The computation can be carried out using the existing R-codes for linear envelope model, with its input replaced by features extracted by the reproducing kernel. We will also present the asymptotic distribution, the asymptotic efficiency gains over the standard RKHS regression, statistical inference, consistent order determination, and confidence and prediction intervals. We compare our method with several existing regression estimators such as the standard RKHS, Principal Component Regression, and Partial Least Squares, which shows significant improvement in estimation efficiency. We apply the method to a chemometrics data involving polymerization reaction.

Asymptotic distribution for partial least square prediction when the number of sample is small

♦ *Liliana Forzani*¹ and *R. Dennis Cook*²

¹Universidad Nacional del Litoral

²University of Minnesota

liliana.forzani@gmail.com

We study consistency and asymptotic behavior of predictions from partial least squares (PLS) regression as the sample size and number of predictors diverge in various alignments.

A unified framework to high dimensional sufficient dimension reduction

♦ *Shanshan Ding*¹, *Wei Qian*¹ and *Lan Wang*²

¹University of Delaware

²University of Miami

sding@udel.edu

Sufficient dimension reduction (SDR) is known to be a powerful tool for achieving data reduction and visualization in regression and classification problems. In this work, we study high dimensional SDR problems including a unified framework of SDR to high-dimensional survival analysis under weak modeling assumptions. This framework includes many popular survival regression models as special cases, and produces a number of practically useful outputs with theoretical guarantees, including a uniformly consistent Kaplan-Meier type estimator of the conditional distribution function of the survival time. Promising applications are demonstrated through simulations and real data analysis on biomedical studies.

Envelope-based Partial Least Squares with Application to Cytokine-based Biomarker Analysis for COVID-19

♦ *Yeonhee Park*¹, *Zhihua Su*² and *Dongjun Chung*³

¹University of Wisconsin

²University of Florida

³Ohio State University

ypark56@wisc.edu

Partial least squares (PLS) regression is a popular alternative to ordinary least squares regression because of its superior prediction performance demonstrated in many cases. In various contemporary applications, the predictors include both continuous and categorical variables. A common practice in PLS regression is to treat the categorical variable as continuous. However, studies find that this practice may lead to biased estimates and invalid inferences (schuberth et al. 2018). Based on a connection between the en-

velope model and PLS, we develop an envelope-based partial PLS estimator that considers the PLS regression on the conditional distributions of the response(s) and continuous predictors on the categorical predictors. Root-n consistency and asymptotic normality are established for this estimator. Numerical study shows that this approach can achieve more efficiency gains in estimation and produce better predictions. The method is applied for the identification of cytokine-based biomarkers for COVID-19 patients, which reveals the association between the cytokine-based biomarkers and patients' clinical information including disease status at admission and demographic characteristics. The efficient estimation leads to a clear scientific interpretation of the results.

Session 5C : Precision Digital Health Care Via Machine Learning

Designing Reinforcement Learning Algorithms for Digital Interventions: Pre-implementation Guidelines

♦ *Anna L. Trella*¹, *Kelly W. Zhang*¹, *Inbal Nahum-Shani*², *Vivek Shetty*³, *Finale Doshi-Velez*¹ and *Susan A. Murphy*¹

¹Harvard University

²University of Michigan

³University of California, Los Angeles

annatrella@g.harvard.edu

Online reinforcement learning (RL) algorithms are increasingly used to personalize digital interventions in the fields of mobile health and online education. Common challenges in designing and testing an RL algorithm in these settings include ensuring the RL algorithm can learn and run stably under real-time constraints, and accounting for the complexity of the environment, e.g., a lack of accurate mechanistic models for the user dynamics. To guide how one can tackle these challenges, we extend the PCS (Predictability, Computability, Stability) framework, a data science framework that incorporates best practices from machine learning and statistics in supervised learning (Yu and Kumbier, 2020), to the design of RL algorithms for the digital interventions setting. Further, we provide guidelines on how to design simulation environments, a crucial tool for evaluating RL candidate algorithms using the PCS framework. We illustrate the use of the PCS framework for designing an RL algorithm for Oralytics, a mobile health study aiming to improve users' tooth-brushing behaviors through the personalized delivery of intervention messages. Oralytics will go into the field in late 2022.

Oblique random survival forests version 2.0: faster and more interpretable

♦ *Byron Jaeger* and *Nicholas Pajewski*

Wake Forest School of Medicine

bjjaeger@wakehealth.edu

The oblique random survival forest (ORSF), developed in 2017 has demonstrated high prediction accuracy in medical applications. For example, a study in 2020 compared the accuracy of several methods to predict the risk of incident heart failure, and ORSF was the top-ranked predictor versus other machine learning algorithms as well as established heart failure risk prediction equations. Although ORSF has exceptional prediction accuracy, its original software (obliqueRSF) is slow. Also, due to its use of oblique recursive partitioning, ORSF models are hard to interpret. We have investigated the mechanics of ORSF and developed two adaptations to address both of these issues. To increase the computational speed of ORSF, we develop a partial Newton Raphson scoring technique to

identify coefficients in linear combinations of variables used to split decision nodes. For interpretation of ORSF, we introduce 'negation importance'. As linear combinations of variables are used in ORSF, a method to estimate variable importance for ORSF should account for the magnitude of a variable's coefficients. Negation importance measures variable importance as the increase in prediction error when all coefficients for a variable are multiplied by -1. As the magnitude of a coefficient in a decision node increases, so does the probability that negating the coefficient will change the node's decisions. We evaluate the speed of the accelerated ORSF algorithm versus leading software packages for standard random survival forests versus the original obliqueRSF software, showing that the accelerated ORSF algorithm is over 750 times faster than its predecessor. We evaluate negation importance by assessing its discrimination of relevant versus irrelevant variables in a simulation study. Both the accelerated ORSF algorithm and negation importance are available in the 'aorsf' R package.

Going Beyond Spike-and-slab: L1-ball Sparsity Prior With Applications On Image Data Analysis

♦ *Leo Duan and Maoran Xu*

University of Florida
li.duan@ufl.edu

Spike-and-slab priors have been viewed as the "gold standard" for variable selection in high dimensional regression. Although some optimality has been shown for signal recovery when both n and p diverge, in practice under a large but finite p (with $p \ll n$), we found that the shrinkage effect of spike-and-slab is often overly-aggressive, causing elimination of true signals in estimates. In this study, we conduct analysis on the alternative, "L1-ball prior", whose sparsity is induced by projecting a continuous precursor random variable onto the surface of L1-ball. We show that when the precursor is allowed to be distributed anisotropically, the induced L1-ball prior gains substantial adaptiveness to different scales of signals. This improvement leads to a much superior signal detection limit, compared to common spike-and-slab priors using iid Bernoulli inclusions. We demonstrate practical advantages of these L1-ball priors in both simulations and data applications, where we obtain a much-improved balance between controlling false discovery rate and maintaining high statistical power. We will demonstrate useful applications of our approach on image data analysis, in which common Bayesian shrinkage priors face tremendous challenges in model specification and/or posterior computation.

Session 5D : Statistical Methods For Complex And High Dimensional Data

Consistent and scalable Bayesian joint variable and graph selection for disease diagnosis leveraging functional brain network

♦ *Xuan Cao¹ and Kyoungjae Lee²*

¹University of Cincinnati

²Sungkyunkwan University
caox4@ucmail.uc.edu

We consider the joint inference of regression coefficients and the inverse covariance matrix for covariates in high-dimensional probit regression, where the predictors are both relevant to the binary response and functionally related to one another. A hierarchical model with spike and slab priors over regression coefficients and the elements in the inverse covariance matrix is employed to simultaneously perform variable and graph selection. We establish joint selection consistency for both the variable and the underlying

graph when the dimension of predictors is allowed to grow much larger than the sample size, which is the first theoretical result in the Bayesian literature. A scalable Gibbs sampler is derived that performs better in high-dimensional simulation studies compared with other state-of-art methods. We illustrate the practical impact and utilities of the proposed method via a functional MRI dataset, where both the regions of interest with altered functional activities and the underlying functional brain network are inferred and integrated together for stratifying disease risk.

Bayesian mixture models, non-local prior formulations and MCMC algorithms

Jairo Alberto Fuquenepatino

UC Davis
jafuquenepatino@ucdavis.edu

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.

Two-component Gibbs samplers: Convergence rate and asymptotic variance

♦ *Qian Qin¹ and Galin Jones²*

¹University of Minnesota

²University of Minnesota
qqin@umn.edu

I will conduct a comparison between deterministic- and random-scan Gibbs samplers in the two-component case. It is found that in terms of asymptotic variance, the random-scan version could greatly outperform the deterministic-scan version in some scenarios, while never being much worse. This is especially the case when there is a large discrepancy between the computational costs of updating the two components. On the other hand, the deterministic-scan version is always superior in terms of convergence rate.

Efficient Algorithms and Theory for High-Dimensional Bayesian Varying Coefficient Models

Ray Bai

University of South Carolina
rbai@mailbox.sc.edu

Nonparametric varying coefficient (NVC) models are useful for modeling time-varying effects on responses that are measured repeatedly. In this talk, we introduce the nonparametric varying coefficient spike-and-slab lasso (NVC-SSL) for Bayesian estimation and variable selection in NVC models. The NVC-SSL simultaneously selects and estimates the significant varying coefficients, while also accounting for temporal correlations. Our model can be implemented using a computationally efficient expectation-maximization (EM) algorithm. We also employ a simple method to make our model robust to misspecification of the temporal correlation structure. In contrast to frequentist approaches, little is known about the large-sample properties for Bayesian NVC models when the dimension of the covariates p grows much faster than sample size n . In this paper, we derive posterior contraction rates for the NVC-SSL model when $p \ll n$ under both correct specification and misspecification of the temporal correlation structure. Thus, our results are derived under weaker assumptions than those seen in other high-dimensional NVC models which assume independent and identically distributed (iid) random errors. Finally, we illustrate our methodology through simulation studies and data analysis. This is joint work with Dr. Mary R. Boland and Dr. Yong Chen.

Session 5E : Modern Streaming Data Analysis: Change-Point Problems And Applications

Detection of multiple change points in multiple profiles

♦ *Jie Chen*¹ and *Shirong Deng*²

¹Augusta University

²Wuhan University
jiechen@augusta.edu

We investigate the problem of detecting boundaries of CNV regions using the DNA-sequencing data from multiple subjects with the consideration of existence of correlation within a subject's sequencing reads. We propose to use the framework of a fused Lasso latent feature model to solve the problem, and experiment with a modified information criterion for selecting the tuning parameter when search for common CNVs shared by multiple subjects. Simulation studies and application on sequencing data show that the proposed approach can effectively identify both common CNVs for multiple subjects and individual CNVs of a single subject profile.

Changepoint Analysis of Hourly Sky-cloudiness Conditions in Canada

*Mo Li*¹, ♦ *Qiqi Lu*¹ and *Xiaolan Wang*²

¹Virginia Commonwealth University

²Environment and Climate Change Canada
qlu2@vcu.edu

Changepoint analysis of non-stationary ordinal categorical time series is always of interest in climate studies. For instance, the sky-cloudiness (or cloud cover) condition in Canada is reported hourly in terms of tenths of the sky dome covered by clouds and hence has 11 ordinal categories. The cloud cover data often contain change-points and exhibit a temporal trend, seasonality, and serial correlation in nature. To properly take into account these features, a likelihood ratio-type statistic is proposed in this talk to test for a single changepoint in hourly categorical time series using a marginalized transition model. This model allows the likelihood-based inference and the series dependence is specified via a first-order Markov chain. An application of our method is illustrated using the hourly sky-cloudiness conditions at Fort St. John Airport in Canada.

Learning under concept drift

Yuekai Sun

University of Michigan
yuekai@umich.edu

In transfer learning and domain adaptation, a learner leverages auxiliary data from a data-rich source domain to improve prediction on a data-poor target domain. We focus on classification problems in which the Bayes decision boundary changes between the source and target domain, making transfer (learning) necessary for optimal performance in the target domain. We characterize the minimax risk of learning under a simple model of posterior drift and develop practical methods that attain the minimax risk.

Inference for Gaussian Multiple Change-point Model via Bayesian Information Criterion

♦ *Yue Niu*¹, *Ning Hao*¹ and *Han Xiao*²

¹University of Arizona

²Rutgers University
yueniu@math.arizona.edu

For a change-point model with a piecewise constant mean structure and additive Gaussian noises, a fundamental inference problem is to determine the existence of change points. Early works usually assume that there is at most one change point. Many recent works can handle multiple changes, however, mainly focus on identifying

individual change points. In particular, it is not clear the weakest condition to guarantee the existence of an asymptotically powerful test. In this talk, we answer this question via a Bayesian information criterion approach.

Session 5F : Emerging Development In The Analysis Of Data With Complex Features

Feature Screening with Large Scale and High Dimensional Survival Data

*Grace Yi*¹, ♦ *Wenqing He*¹ and *Raymond Carroll*²

¹University of Western Ontario

²Texas A&M University, University of Technology Sydney
whe@stats.uwo.ca

Data with a huge size present great challenges in modelling, inferences and computation. In handling big data, much attention has been directed to settings with "large p small n", and relatively less work has been done to address problems with p and n being both large, though data with such a feature have now become more accessible than before, where p represent the number of variables and n stands for the sample size. The big volume of data does not automatically ensure good quality of inferences because a large number of unimportant variables may be collected in the process of gathering informative variables. To carry out valid statistical analysis, it is imperative to screen out noisy variables that have no predictive value for explaining the outcome variable. In this paper, we develop a screening method of handling large sized survival data, where the sample size n is large and the dimension p of covariates is of non-polynomial order of the sample size n, or the so called NP-dimension. We rigorously establish theoretical results for the proposed method and conduct numerical studies to assess its performance. Our research offers multiple extensions of existing work and enlarges the scope of high dimensional data analysis. The proposed method capitalizes on the connections among useful regression settings and offers a computationally efficient screening procedure. Our method can be applied to different situations with large scale data including genomic data analysis.

Analysis of the Cox Model with Longitudinal Covariates with Measurement Errors and Partly Interval Censored Failure Times, with Application to an AIDS Clinical Trial

♦ *Yanqing Sun*¹, *Qingning Zhou*¹ and *Peter Gilbert*²

¹University of North Carolina at Charlotte

²Fred Hutchinson Cancer Research Center and University of Washington
yasun@uncc.edu

Time-dependent covariates are often measured intermittently and with measurement errors. Motivated by the AIDS Clinical Trials Group (ACTG) 175 trial, this paper develops statistical inferences for the Cox model for partly interval censored failure times and longitudinal covariates with measurement errors. The conditional score methods developed for the Cox model with measurement errors and right censored data are no longer applicable to interval censored data. Assuming an additive measurement error model for a longitudinal covariate, we propose a nonparametric maximum likelihood estimation approach by deriving the measurement error induced hazard model that shows the attenuating effect of using the plug-in estimate for the true underlying longitudinal covariate. An EM algorithm is devised to facilitate maximum likelihood estimation that accounts for the partly interval censored failure times. Simulation studies show that the proposed methods perform well with

satisfactory finite-sample performances and that the naive methods ignoring measurement error or using the plug-in estimate can yield large biases. A hypothesis testing procedure for the measurement error model is proposed. The proposed methods are applied to the ACTG 175 trial to assess the associations of treatment arm and time-dependent CD4 cell counts on the composite clinical endpoint of AIDS or death.

Learning Optimal Dynamic Treatment Regimens Subject to Stagewise Risk Control

Mochuan Liu¹, Yuanjia Wang², Haoda Fu³ and [◆]Donglin Zeng¹

¹University of North Carolina

²Columbia University

³Eli Lilly and Company
dzeng@bios.unc.edu

Dynamic treatment regimes (DTRs) aim at tailoring individualized sequential treatment rules that maximize cumulative beneficial outcomes by accommodating patient's heterogeneity into decision making. For many chronic diseases including type 2 diabetes mellitus (T2D), treatments are usually multifaceted in the sense that the aggressive treatments with higher expected reward are also likely to elevate the risk of acute adverse events. In this paper, we propose a new weighted learning framework, namely benefit-risk dynamic treatment regimes (BR-DTRs), to address the benefit-risk trade-off. The new framework relies on a backward learning procedure by restricting the induced risk of the treatment rule to be no larger than a pre-specified risk constraint at each treatment stage. Computationally, the estimated treatment rule solves a weighted support vector machine problem with a modified smooth constraint. Theoretically, we show that the proposed DTRs are Fisher consistent and we further obtain the convergence rates for both the value and risk functions. Finally, the performance of the proposed method is demonstrated via extensive simulation studies and application to a real study for T2D patients.

A new Bayesian method for handling covariate measurement error and detection limit in regression models

[◆]Muhire Kwizera¹, Roderick Little², Matthew Perzanowski³ and Qixuan Chen¹

¹Department of Biostatistics, Columbia University

²Department of Biostatistics, University of Michigan

³Department of Environmental Health Sciences, Columbia University
mhk2159@cumc.columbia.edu

Measurement error and limit of detection (LOD) are common problems in environmental health studies that should be properly accounted for as they can bias regression results. We propose a new Bayesian method, based on the general location model, to correct for covariate measurement error and to handle data below LOD in unadjusted and adjusted regression models. We consider the external calibration setting where in addition to the main study sample of interest, another calibration sample providing information only on the measurement error is available. Unlike conventional methods that only use the calibration data for model estimation, our method uses observed data from both the calibration and main study samples and thus incorporates relationships between model variables in measurement error adjustment. We use the nondifferential measurement error (NDME) assumption to identify all parameters in our model, and we use the sweep operator to estimate regression parameters of interest. In simulations, our method yields reduced bias, smaller mean squared error, and with credible interval coverage closer to the nominal level compared to existing methods. The

improvement becomes more pronounced with increased measurement error, correlation between covariates, and a stronger covariate effect. We applied the new method to the Neighborhood Asthma and Allergy Study (NAAS) to examine the association between indoor allergen concentrations and asthma morbidity among urban children.

Session 5G : Statistical Leadership In Drug Development In The New Era Of Data Science

Opportunities and Challenges of Using Real-world Data for Signal Identification and Evidence Generation to Inform Study Design and Scientific Questions in Medical Research

Yiyue Lou

Vertex Pharmaceuticals
yiyue.lou@vrtx.com

The rapid increase in the volume, type, and accessibility of real-world data has been identified as a great opportunity for today's statisticians. The complexity of real-world data generates unique challenges in identifying the appropriate data sources, designing investigations, and analyzing and interpreting findings from such data, resulting in an increasing need for statistical skills. We introduce different sources of real-world data including disease registry, electronic health record, medical claims, and medical chart review. We discuss their advantages and limitations, along with practical examples of how such data can be used for signal detection and evidence generation to inform study design and scientific questions, as well as how to enhance statistical leadership in medical research using such real-world data.

Empowering Real-World Evidence Generation in Rare Conditions: Collaborative data initiatives

[◆]Jia Zhong, James Signorovitch and Eric Wu

Analysis Group
Jia.Zhong@analysisgroup.com

Major strides have been made in improving the availability and usability of real-world data (RWD). However, the generation of real-world evidence (RWE) is a multifaceted process facing many challenges globally, especially in rare conditions. This presentation will analyze the key issues that hinder the generation of high quality RWD, present several case studies in the US and China, and discuss creative solutions and the value of collaborative data initiatives. The first case study will discuss opportunities and approaches to overcome data access barriers and generate early RWE in support of market access in China. The second case study will follow to share recent advances in data quality in China's National Longitudinal Cohort of Hematological Diseases (NICHE), in which the researchers developed and validated algorithms to generate high-quality RWD in order to study complex patient journey in a rare hematologic condition. We will then feature case studies from the US to shed light on design and methodology considerations, which are important to mitigate potential bias in the creation of historical control arms and ensure consistency across heterogeneous data sources. This presentation aims to introduce these recent creative initiatives in generating and applying RWD to the audience, and stimulate discussions to further advancing the RWE research and eventually fostering a sustainable and scalable RWD ecosystem.

Assessing Mediation Processes using Joint Longitudinal Models in the Framework of Individual Measurement Occasions

[◆]Jin Liu¹, Robert Perera² and Yijie Zhou¹

¹Vertex Pharmaceuticals

²Virginia Commonwealth University
Jin.Liu@vrtx.com

In clinical trials with a longitudinal design, the primary endpoint and multiple secondary endpoints are often recorded repeatedly. Traditionally, these endpoints are analyzed separately using the mixed model for repeated measures (MMRM) if they are continuous. However, separate MMRM fails to capture the relationship between these longitudinal outcomes, and a joint modeling approach is desired. In addition, clinical indicators often improve first in clinical trials, leading to improvement in patient-reported outcomes related to illness burden or health-related quality of life. Therefore, the joint modeling approach should allow different longitudinal patterns for each outcome. This research develops two longitudinal mediation models to address the above challenges. We define the mediational process as either the baseline covariate or the change in the covariate influencing the change of the mediator, which, in turn, affects the change of the outcome. The developed models utilize the linear-linear piecewise functional form to mimic short- and long-term recovery periods. The proposed models are evaluated using simulation studies. Our simulation studies demonstrate that the proposed mediational models can provide unbiased and accurate point estimates with target coverage probabilities. We also demonstrate the proposed models through a real-world data analysis.

Discussion

Yijie Zhou

Vertex
yijie.zhou@vrtx.com

Discussion of the three invited talks in this session.

Session 5H : Student Paper Competition Winners

Sensitivity Analysis under the f-Sensitivity Models: A Distributionally Robust Optimization Viewpoint

◆ *Ying Jin¹, Zhimei Ren² and Zhengyuan Zhou³*

¹Stanford University

²University of Chicago

³New York University
ying531@stanford.edu

The growing availability of observational data in various application regimes provides exciting opportunities for causal inference, especially in situations where randomization can be costly or unethical. However, observational data face the key challenge of unmeasured confounding: there might be unobserved variables that influences both the treatment mechanism and the outcomes, invalidating the causal conclusions inferred from the data. While such violation is not verifiable, sensitivity analysis could help gauge the robustness and credibility of the causal conclusions.

In this paper, we propose a new class of sensitivity models where the selection bias caused by unmeasured confounding factors is bounded on average. It generalizes commonly adopted uniform bound on selection bias in the literature, and might be more suitable for situations where the selection bias grows unbounded locally but, due to the impact of such region being small overall, is controlled at the population level. The new class of models motivates an illuminating perspective from distributionally robust optimization for sensitivity analysis: we show that the counterfactual distribution, the essential for inferring treatment effects, is within certain distance from the observed distribution. Based on this perspective, we represent the bounds on counterfactual means via distributionally robust optimization programs. We then design procedures to estimate these bounds, and show that our procedure is doubly robust

to the estimation error of nuisance components and remains valid even when the optimization step is off. In addition, we establish Wald-type inference guarantee that is again robust to the optimization step. We demonstrate our method and verify its validity with numerical experiments.

Fast Distributed Principal Component Analysis for Large-Scale Federated Data

◆ *Shuting Shen, Junwei Lu and Xihong Lin*

Harvard University
shs145@g.harvard.edu

Principal component analysis (PCA) is one of the most popular methods for dimension reduction. In light of the rapidly increasing large-scale data in federated ecosystems, the traditional PCA method is often not applicable due to privacy protection considerations and large computational burden. Fast PCA algorithms have been proposed to lower the computational cost but cannot handle federated data. Distributed PCA algorithms have been developed to handle federated data but are not computationally efficient when data at each site are very large. In this paper, we propose the FAST Distributed (FADI) PCA method which applies fast PCA to site-specific data using multiple random sketches and aggregates the results across sites. We perform a non-asymptotic theoretical study to show that FADI enjoys the same error rate as the traditional full sample PCA and a much smaller order of computational burden compared to existing methods. We perform extensive simulation studies and show that FADI substantially outperforms the other methods in computational efficiency without sacrificing statistical accuracy. We apply FADI to the analysis of the 1000 Genomes data to study the population structure.

High-Dimensional Dynamic Process Monitoring By PCA-Based Sequential Learning

◆ *Xiulin Xie and Peihua Qiu*

University of Florida
xiulin.xie@ufl.edu

Sequential process monitoring has broad applications. In practice, process characteristics to monitor often have a high dimensionality, partly due to the fast progress in data acquisition techniques. Thus, statistical process control (SPC) research for monitoring high-dimensional processes is in rapid development in recent years. Most existing SPC charts for monitoring high-dimensional processes are designed for conventional cases in which the in-control (IC) process observations at different time points are assumed to be independent and identically distributed. In practice, however, serial correlation almost always exists in the observed sequential data, and the longitudinal pattern of the process to monitor could be dynamic in the sense that its IC distribution would vary over time (e.g., seasonality). In this paper, we develop a novel SPC chart for monitoring high-dimensional dynamic processes. The new method is based on nonparametric longitudinal modeling for describing the longitudinal pattern of the process under monitoring, principal component analysis for dimension reduction, and a sequential learning algorithm for developing an effective decision rule. It can well accommodate time-varying IC process distribution, serial data correlation, and nonparametric data distribution. The proposed method has been shown effective for air pollution surveillance.

Supervised Learning of Physical Activity Features from Functional Accelerometer Data

◆ *Margaret Banker and Peter X.K. Song*

University of Michigan
mbanker@umich.edu

Accelerometry data enables scientists to extract personal digital features that can benefit precision health decision making. Existing methods in accelerometry data analysis typically begin with discretizing summary single-axis counts by certain fixed cutoffs into several activity categories, such as Vigorous, Moderate, Light, and Sedentary. One well-known limitation is that the chosen cutoffs have often been validated with restricted settings, and thus they cannot be generalizable across populations, devices, or studies. In this paper, we develop a data-driven approach to overcome this bottleneck in the analysis of activity data, in which we holistically summarize a subject's activity profile using Occupation-Time curves (OTCs). Being a functional predictor, OTC describes the percentage of time spent at or above a continuum of activity count levels. We develop multi-step adaptive learning algorithms to perform a supervised learning via a scale-functional regression model that contains OTC as the functional predictor of interest as well as other covariates. Our learning algorithm first incorporates a hybrid approach of fused lasso for grouping and Hidden Markov Model for change-point detection, and then executes a few refinement learning steps to yield activity windows of interest. We demonstrate good performances of this learning algorithm using simulations as well as real world data analysis to assess the influence of physical activity on biological aging.

Session 6A : Recent Advances In Mendelian Randomization

Inference of nonlinear causal effects with GWAS summary data

◆ Ben Dai¹, Chunlin Li², Haoran Xue², Wei Pan² and Xiaotong Shen²

¹The Chinese University of Hong Kong

²The University of Minnesota
bendai@cuhk.edu.hk

Large-scale genome-wide association studies (GWAS) have offered an exciting opportunity to discover putative causal genes or risk factors associated with diseases by using SNPs as instrumental variables (IVs). However, conventional approaches assume linear causal relations partly for simplicity and partly for the only availability of GWAS summary data. In this work, we propose a novel model to incorporate nonlinear relationships across IVs, an exposure, and an outcome, which is robust against violations of the valid IV assumptions and permits the utilization of GWAS summary data. We decouple the estimation of a marginal causal effect and a nonlinear causal transformation, where the former is estimated via sliced inverse regression and a sparse instrumental variable regression, and the latter is estimated by a ratio-adjusted inverse regression. On this ground, we propose an inferential procedure. An application of the proposed method to the ADNI gene expression data and the IGAP GWAS summary data identifies 17 causal genes associated with Alzheimer's disease, including APOE and TOMM40, in addition to 7 other genes missed by two-stage least squares considering only linear relationships. Our findings suggest that nonlinear modeling is required to unleash the power of IV regression for identifying potentially nonlinear gene-trait associations. Accompanying this paper is our Python library nl-causal that implements the proposed method.

Causal analysis with rerandomization estimators (CARE)

◆ Chong Wu¹ and Jingshen Wang²

¹FLORIDA STATE UNIVERSITY

²University of California, Berkeley

cwu3@fsu.edu

Mendelian randomization (MR) studies use genetic variants associated with modifiable exposures as instrumental variables to assess their possible causal relationship with outcomes in observational studies. In recent years, the number of published two-sample MR studies has increased rapidly, partially due to the increasing availability of large-scale genome-wide association studies (GWAS) summary data. These two-sample MR studies often employ the same sample to select relevant genetic variants and construct final causal estimates, leading to biased causal effect estimates due to the winner's curse phenomenon. On the other hand, the validity of MR analyses critically depends on instrumental variable assumptions, which often be violated in applications due to widespread pleiotropic effects. Also, there are maybe unknown sample overlaps between the two GWAS summary datasets due to the current trends in collecting biobank datasets. To fill these gaps, we propose a new approach, referred to as Causal Analysis with Rerandomized Estimator (CARE), that corrects winner's curse bias, weak pleiotropic effects, and unknown sample overlaps. We demonstrate the utilities of CARE through negative control analyses and its applications to identify possible causal risk factors for COVID-19 severity. This is joint work with Jingshen Wang.

Breaking the Winner's Curse in Mendelian Randomization: Rerandomized Inverse Variance Weighted Estimator

Xinwei Ma¹, ◆ Jingshen Wang² and Chong Wu³

¹UC San Diego

²UC Berkeley

³Florida State University
jingshenwang@berkeley.edu

Developments in genome-wide association studies and the increasing availability of summary genetic association data have made the application of two-sample Mendelian Randomization (MR) with summary data increasingly popular. Conventional two-sample MR methods often employ the same sample for selecting relevant genetic variants and for constructing final causal estimates. Such a practice often leads to biased causal effect estimates due to the well known "winner's curse" phenomenon. To address this fundamental challenge, we first examine the consequence of winner's curse on causal effect estimation both theoretically and empirically. We then propose a novel framework that systematically breaks the winner's curse, leading to unbiased association effect estimates for the selected genetic variants. Building upon the proposed framework, we introduce a novel rerandomized inverse variance weighted estimator that is consistent when selection and parameter estimation are conducted on the same sample. Under appropriate conditions, we show that the proposed RIVW estimator for the causal effect converges to a normal distribution asymptotically and its variance can be well estimated. We illustrate the finite-sample performance of our approach through Monte Carlo experiments and two empirical examples.

Constrained maximum likelihood-based Mendelian randomization robust to both correlated and uncorrelated pleiotropic effects

◆ Haoran Xue¹, Xiaotong Shen² and Wei Pan¹

¹Division of Biostatistics, School of Public Health, University of Minnesota

²School of Statistics, University of Minnesota
xuexx268@umn.edu

With the increasing availability of large-scale GWAS summary data on various complex traits and diseases, there have been tremendous

interests in applications of Mendelian randomization (MR) to investigate causal relationships between pairs of traits using SNPs as instrumental variables (IVs) based on observational data. In spite of the potential significance of such applications, the validity of their causal conclusions critically depends on some strong modeling assumptions required by MR, which may be violated due to the widespread (horizontal) pleiotropy. Although many MR methods have been proposed recently to relax the assumptions by mainly dealing with uncorrelated pleiotropy, only a few can handle correlated pleiotropy, in which some SNPs/IVs may be associated with hidden confounders, such as some heritable factors shared by both traits. Here we propose a simple and effective approach based on constrained maximum likelihood and model averaging, called cMLMA, applicable to GWAS summary data. To deal with more challenging situations with many invalid IVs with only weak pleiotropic effects, we modify and improve it with data perturbation. Extensive simulations demonstrated that the proposed methods could control the type I error rate better while achieving higher power than other competitors. Applications to 48 risk factor-disease pairs based on large-scale GWAS summary data of 3 cardio-metabolic diseases (coronary artery disease, stroke, and type 2 diabetes), asthma, and 12 risk factors confirmed its superior performance.

Session 6B : Recent Advances In Dimension Reduction Techniques

Significance testing for canonical correlation analysis in high dimensions

Ian Mckeague¹ and [◆]Xin Zhang²

¹Columbia University

²Florida State University
henry@stat.fsu.edu

We consider the problem of testing for the presence of linear relationships between large sets of random variables based on a post-selection inference approach to canonical correlation analysis. The challenge is to adjust for the selection of subsets of variables having linear combinations with maximal sample correlation. To this end, we construct a stabilized one-step estimator of the euclidean-norm of the canonical correlations maximized over subsets of variables of pre-specified cardinality. This estimator is shown to be consistent for its target parameter and asymptotically normal provided the dimensions of the variables do not grow too quickly with sample size. We also develop a greedy search algorithm to accurately compute the estimator, leading to a computationally tractable omnibus test for the global null hypothesis that there are no linear relationships between any subsets of variables having the pre-specified cardinality. Further, we develop a confidence interval for the target parameter that takes the variable selection into account.

Dimension Reduction Forests: Local Variable Importance using Structured Random Forests

[◆]*Joshua Loyal¹, Ruoqing Zhu¹, Yifan Cui² and Xin Zhang³*

¹University of Illinois at Urbana-Champaign

²National University of Singapore

³Florida State University
jloyal2@illinois.edu

Random forests are one of the most popular machine learning methods due to their accuracy and variable importance assessment. However, random forests only provide variable importance in a global sense. There is an increasing need for such assessments at a local

level, motivated by applications in personalized medicine, policy-making, and bioinformatics. We propose a new nonparametric estimator that pairs the flexible random forest kernel with local sufficient dimension reduction to adapt to a regression function's local structure. This allows us to estimate a meaningful directional local variable importance measure at each prediction point. We develop a computationally efficient fitting procedure and provide sufficient conditions for the recovery of the splitting directions. We demonstrate significant accuracy gains of our proposed estimator over competing methods on simulated and real regression problems. Finally, we apply the proposed method to seasonal particulate matter concentration data collected in Beijing, China, which yields meaningful local importance measures. The methods presented here are available in the drforest Python package.

Envelope model for function-on-function linear regression

[◆]*Zhihua Su¹, Bing Li² and Dennis Cook³*

¹University of Florida

²Pennsylvania State University

³University of Minnesota
zhihuasu@stat.ufl.edu

The envelope model is a recently developed methodology for multivariate analysis that enhances estimation accuracy by exploiting the relation between the mean and eigenstructure of the covariance matrix. We extend the envelope model to function-on-function linear regression, where the response and the predictor are assumed to be random functions in Hilbert spaces. We use a double envelope structure to accommodate the eigenstructures of the covariance operators for both the predictor and the response. The central idea is to establish a one-to-one relation between the functional envelope model and the multivariate envelope model and estimate the latter by existing method. We also developed the asymptotic theories, confidence and prediction bands, an order determination method along with its consistency, and a characterization of the efficiency gain by the proposed model. Simulation comparisons with the standard function-on-function regression and data applications show significant improvement by our method in terms of cross-validated prediction error.

Session 6C : Statistical Methods For Assessing Genomic Heterogeneity

Robust Statistical Inference for Cell Type Deconvolution

[◆]*Jingshu Wang and Dongyue Xie*

University of Chicago
jingshuw@uchicago.edu

Cell type deconvolution is a computational approach to infer proportions of individual cell types from bulk transcriptomics data. Though many new methods have been developed for cell type deconvolution, most of them only provide point estimation of the cell type proportions. On the other hand, estimates of the cell type proportions can be very noisy due to various sources of bias and randomness, and ignoring their uncertainty may greatly affect the validity of downstream analyses. In this paper, we propose a comprehensive statistical framework for cell type deconvolution and construct asymptotically valid confidence intervals both for each individual's cell type proportion and for quantifying how cell type proportions change across multiple bulk individuals in downstream regression analyses. Our analysis takes into account various factors including the biological randomness of gene expressions across cells and individuals, gene-gene dependence, and the cross-platform biases

and sequencing errors, and avoids any parametric assumptions on the data distributions. We also provide identification conditions of the cell type proportions when there are arbitrary platforms-specific biases across sequencing technologies.

Single-cell eco-evolutionary dynamics of intratumor heterogeneity

Meghan Ferrall-Fairbanks

University of Florida
mferrall.fairbanks@bme.ufl.edu

Researchers have recognized that a one-size-fits-all approach is not effective at treating cancer and that tumor heterogeneity plays an important role in response. Current dogma stipulates that this heterogeneity results from compounding genetic and epigenetic changes and instability, ultimately driving unfavorable outcomes for these patients. Nonetheless, some cancers, including many pediatric cancers and some leukemias, have limited genomic diversity. As a result, we have a limited ability to stratify patients into high versus low-risk groups. To address this, we explore intratumor heterogeneity using single-cell transcriptomics data to quantify and identify driving phenotypes in tumor evolution. We leverage the generalized diversity index (GDI) from ecology, which allows us to tailor the scale of cellular diversity in a given context. We show that the order of diversity parameter in GDI allows us to either emphasize clonal richness at low values while high values shift the analysis toward the abundance of potential drivers of the tumor evolution. We have explored these phenotypic changes in both engineered (in vitro) and patient samples using single-cell RNA sequencing. Our analyses show how quantifying intratumor heterogeneity with GDI is a powerful tool to understand eco-evolutionary dynamics of a patient's tumor.

Neural Network Models for Sequence-Based TCR and HLA Association Prediction

♦ *Si Liu, Phil Bradley and Wei Sun*

Fred Hutchinson Cancer Center
sliu3@fredhutch.org

T-cell mediated immunity relies on T-cell receptor (TCR) to recognize peptides bound by major histocompatibility complex (MHC). The human MHC is also called the human leukocyte antigen complex (HLA). It is important to account for HLA information in TCR analysis for predicting infection status. Currently in literature, the association between TCR and HLA is assessed through co-occurrence pattern. In this work we explore the capacity of certain neural network models to predict the association between TCR and HLA, based on the amino acid sequence information of HLA at certain positions and both CDR3 amino acid sequence information and V allele information of TCR. Our model can make predictions on HLA and TCR with amino acid sequences not seen during training.

A statistical framework for cell-type-specific transcriptomics-wide association studies with an application to breast cancer

Xiaoyu Song

Icahn School of Medicine at Mount Sinai
xiaoyu.song@mountsinai.org

Human bulk tissue samples comprise multiple cell types with diverse roles in disease etiology. Conventional transcriptome-wide association study (TWAS) approaches predict gene expression at the tissue level from genotype data, without considering cell-type heterogeneity, and test associations of the predicted tissue-level gene expression with disease. Here we develop MiXcan, a new TWAS approach that predicts cell-type-specific gene expression levels, identifies disease-associated genes via combination of cell-

type-specific association signals for multiple cell types, and provides insight into the disease-critical cell type. We conducted the first cell-type-specific TWAS of breast cancer in 58,648 women and identified 12 transcriptome-wide significant genes using MiXcan compared with only eight genes using conventional approaches. Importantly, MiXcan identified genes with distinct associations in mammary epithelial versus stromal cells, including three new breast cancer susceptibility genes. These findings demonstrate that cell-type-specific TWAS can reveal new insights into the genetic and cellular etiology of breast cancer and other diseases.

Session 6D : Novel Statistical Modeling And Computing Methods For Complex Data

New Bounded response models for target variables

Jorge Bazan

USP
jlbazan@icmc.usp.br

In this talk we introduce a new formulation to propose new bounded response models for target variables. For the new proposed models, we develop estimation methods and implement it computationally. An application for illustrate the advantages of the propose is presented showing the advantages of our proposal.

Penalized complexity priors for the skewness parameter of power links

♦ *Jose Ordonez¹, Marcos Prates², Jorge Bazan³ and Victor Lachos⁴*

¹Federal University of Bahia

²Federal University of Minas Gerais

³ICMC - USP

⁴University of Connecticut
ordonezjosealejandro@gmail.com

The choice of the prior distribution is a key aspect of the Bayesian method, however in many cases, such as the family of power links, this is not trivial. In this work, we introduce a penalized complexity prior of the skewness parameter for this family, which is useful to deal with imbalanced data. We derive a general expression for this density and show its usefulness for some particular cases such as the power logit and the power probit links. To obtain posterior samples, we use Hamiltonian Monte Carlo, which avoids the random walk behavior of Metropolis and Gibbs sampling algorithms. A simulation study and a real data application are used to assess the efficiency of the introduced densities in comparison with the Gaussian and uniform priors. Results show improvement in point and credible interval estimation for the considered models when using the penalized complexity prior in comparison to other well-known standard priors.

Linear Mixed-effects Models For Censored Data With Serial Correlation Errors Using The Multivariate Student's T-distribution

♦ *Kelin Zhong¹, Rommy C. Olivari², Aldo M. Garay² and Victor H. Lachos³*

¹Department of Statistics, UConn

²Department of Statistics, Federal University of Pernambuco

³Department of Statistics, University of Connecticut
kelin.zhong@uconn.edu

The purpose of this paper is develop a practical framework for the analysis of the linear mixed-effects models for censored data with serial correlation errors, using the multivariate Student's t-distribution (t-LMEC), being a flexible alternative to the use of the

corresponding normal distribution. We propose an efficient ECM algorithm for computing the maximum likelihood estimates for the t-LMEC with standard errors of the fixed effects and likelihood function as a by-product. This algorithm uses closed-form expressions at the Estep, which relies on formulas for the mean and variance of a truncated multivariate Student's t-distribution. In order to illustrate the usefulness of the proposed newly methodology, artificial and real data set are analyzed. The proposed algorithm and methods are implemented in the R package ARpLMEC.

Session 6E : Modern Streaming Data Analysis: Detection And Identification

Low-Rank Robust Subspace Tensor Clustering for Metro Passenger Flow Modeling

Nurretin Sergin, Jiuyun Hu and ♦ Hao Yan

Arizona State University
haoyan@asu.edu

Tensor clustering has become an important topic specifically in spatiotemporal modeling due to its ability to cluster spatial modes and temporal modes (e.g., time of the day or day of the week). Our motivating example is from subway passenger flow modeling where similarities between stations are commonly found. In this presentation, we design a tensor-based subspace clustering and anomaly decomposition technique to achieve simultaneous dimensionality reduction, spatial clustering, and sparse outlier decomposition. This paper combines the Tucker decomposition, sparse anomaly decomposition, and subspace clustering for outlier-robust dimension reduction and clustering for high-dimensional tensors. The effectiveness of the proposed framework is validated through the simulation study and a real case study in the station clustering based on real passenger flow data.

Optimal Parallel Sequential Change Detection under Generalized Performance Measures

Zexian Lu¹, Yunxiao Chen² and ♦ Xiaou Li¹

¹University of Minnesota

²London School of Economics and Political Sciences
lix1766@umn.edu

We consider the detection of change points in parallel data streams, a problem widely encountered when analyzing large-scale real-time streaming data. Each stream may have its own change point, at which its data has a distributional change. With sequentially observed data, a decision maker needs to declare whether changes have already occurred to the streams at each time point. Once a stream is declared to have changed, it is deactivated permanently so that its future data will no longer be collected. This is a compound decision problem in the sense that the decision maker may want to optimize certain compound performance metrics that concern all the streams as a whole. Thus, the decisions are not independent for different streams. Our contribution is three-fold. First, we propose a general framework for compound performance metrics that includes the ones considered in the existing works as special cases and introduces new ones that connect closely with the performance metrics for single-stream sequential change detection and large-scale hypothesis testing. Second, we develop data-driven decision procedures under this framework. Finally, optimality results are established for the proposed decision procedures.

Differentially private approaches for streaming data analysis

Wanrong Zhang

Harvard University

wanrongzhang@fas.harvard.edu

In many common applications of sequential decision-making problems, the data can be highly sensitive and may require privacy protection. For example, sequential hypothesis testing is used in clinical trials, where doctors sequentially collect data from patients and must determine when to stop recruiting patients and whether the treatment is effective; Sequential change-point is used to detect a disease outbreak based on real-time data about hospital visits or detect changes in activity within the home through smart home IoT devices. The field of differential privacy has been developed to offer data analysis tools with strong privacy guarantees. Informally, an algorithm that is ϵ -differentially private ensures that any particular output of the algorithm is at most e^{ϵ} more likely when a single data entry is changed. In the past decade, the theoretical computer science community has developed a wide variety of differentially private algorithms for many statistical tasks. In this talk, I will talk about differentially private approaches to the two fundamental problems for streaming data analysis: change-point detection and sequential hypothesis testing problems. Our algorithms offer strong theoretical privacy guarantees. We also provide theoretical analysis on statistical performance and empirically validate our results.

Active sequential change-point detection under sampling control

Yajun Mei

Georgia Institute of Technology
yme1@isye.gatech.edu

Active sequential change-point detection problem occurs in many real-world problems such smart manufacturing or biosurveillance when one monitors multi-dimensional or high-dimensional data streams under the sampling control due to limited capacity in data acquisition, transmission or processing. In such a scenario, one needs decide how to smartly observe which local components or features of multi-dimensional streaming data at each and every time, and then uses the observed incomplete data to quickly raise an alarm once a change has occurred subject to the false alarm constraint. In this talk, we present two of our latest research by developing efficient active sequential change-point detection algorithms under the sampling control through bandit sampling policies: one is Robbins' win-stay, lose-switch policy that leads to asymptotically optimal algorithm when monitoring low-dimensional data, and the other is Thompson sampling policy that yields efficient scalable schemes for online monitoring high-dimensional data. Numerical simulations and case studies will be presented to demonstrate the usefulness of our proposed algorithms, and future potential research directions will also be discussed.

Session 6F : Deep Learning With Application And Uncertainty Quantification

Random walk with restart with graph embedded neural network to inform potential targets

♦ Yushi Liu, Bochao Jia and Rick Higgs

Eli Lilly

liu-yushi@lilly.com

Target identification has remained as one important task for pharma industry. Given a variety of different data source available (such as expression data, methylation data), meta-analysis could be used to generate robust gene ranking for disease mechanism understanding. Since the current datasets are all heavily transcriptome focused, they

may not match well with the disease mechanism. PPI network could help to enhance the understanding of the disease. To further facilitate drug target discovery, graph embedded deep neural network has been used to shape the generic network to a disease-specific and tissue-specific network combined with random walk with restart, we increased the sensitivity of target identification.

Learning interactions in Reaction Diffusion Equation with Deep Learning

Sichen Chen¹, [◆]Nicolas Brunel², Xin Yang³ and Xinpeng Cui¹

¹Department of Statistics, University of California, Riverside

²Laboratoire de Mathématiques et Modélisation d'Evry, ENSIIE

³Department of Mathematics, University of California, Riverside
nicolas.brunel@ensiie.fr

Nonlinear Reaction-diffusion equations are mathematical models that are extensively used in scientific fields. The question of recovering these PDEs based on experimental data is very important but is still widely open in general situations. Recently, significant advances have been made towards this question by taking advantage of methods from deep learning. In literature, most results are concerning the case in which the nonlinear reaction terms are polynomials of the unknown functions. In this paper, we study more complicated equations where the interactions between species exhibit saturation effect (such as logistic function). We propose to learn them as fractions of polynomials. Such equations often appear in activator-inhibitor systems, such as the Gierer-Meinhardt system and the Thomas system. By combining a the modified PDE-Net method and some sparsity analysis, we manage to discover the hidden terms, in particular the fractional terms, as well as their coefficients in these equations according to the simulated data.

An optimal transport approach for selecting a representative subsample

Ping Ma

University of Georgia
pingma@uga.edu

Subsampling methods aim to select a subsample as a surrogate for the observed sample. Such methods have been used pervasively in large-scale data analytics, active learning, and privacy-preserving analysis in recent decades. Instead of model-based methods, we study model-free subsampling methods, which aim to identify a subsample that is not confined by model assumptions. Existing model-free subsampling methods are usually built upon clustering techniques or kernel tricks. Most of these methods suffer from either a large computational burden or a theoretical weakness. In particular, the theoretical weakness is that the empirical distribution of the selected subsample may not necessarily converge to the population distribution. Such computational and theoretical limitations hinder the broad applicability of model-free subsampling methods in practice. In this talk, I present a novel model-free subsampling method by utilizing optimal transport techniques.

Distribution-free uncertainty quantification for classification

[◆]Sasha Podkopaev and Aaditya Ramdas

Carnegie Mellon University
apodkopa@andrew.cmu.edu

Without additional post-processing, ML models often fail to accurately represent uncertainty, with a tendency to make over-confident predictions. However, confident mispredictions might have disastrous consequences in applications, e.g., medical diagnosis. Supplementing model predictions with the corresponding measures of uncertainty allows an end-user to decide whether to intervene on hard-to-classify examples. Over the recent years, predictive mod-

els have enormously increased in size and complexity, and state-of-the-art accuracy is often achieved after performing various engineering tweaks during training. I will talk about assumption-lean uncertainty quantification when the underlying model is treated as a black box. After considering the iid case, I will focus on settings where models encounter variability in the input distribution of the test data. I will discuss appropriate corrections that allow handling several forms of distribution shifts without retraining the model.

Session 6G : Recent Advances In Clinical Trial Design And Practice

A hybrid efficacy/effectiveness estimand for binary composite endpoints in clinical trials

[◆]Xingyuan Li and Nathan Morris

Eli Lilly and Company
lixingyuan@lilly.com

The ICH E9 (R1) guidance highlighted the need for more clarity in defining the estimand at trial planning, including how to handle intercurrent events to align with the current treatment goal within the therapeutic area. However, current practice using specific statistical analyses oftentimes fails to take into consideration the reason for intercurrent events, resulting in lack of interpretability to the underlying clinical question. Motivated from clinical trials in Ulcerative Colitis, we proposed a hybrid efficacy/effectiveness (HEE) estimand that can be used to answer a shared question of interest to patients, physicians, and the regulators: What is the treatment difference comparing Trt A versus Trt B in response rate, had all patients continued on treatment (i.e., taken as labeled) except those who discontinued due to lack of efficacy (i.e., no treatment benefit) or discontinued due to adverse events (i.e., intolerance of the drug), regardless of use of other conventional UC therapies (i.e., reflect clinical practice when deemed necessary)? A simulation study was conducted to illustrate and compare several estimators of the HEE estimand, when the primary target of estimation is a binary composite endpoint with ordinal scales for the subscores. When imputing missing ordinal variables due to administrative reasons using multiple imputation under the multivariate normal assumption, a calibrated rounding technique showed advantage over simple rounding.

Bayesian adaptive model selection design for optimal biological dose finding in phase I/II clinical trials

Ruitao Lin

The University of Texas MD Anderson Cancer Center
RLin@mdanderson.org

Identification of the optimal dose presents a major challenge in drug development with molecularly targeted agents, immunotherapy, as well as chimeric antigen receptor T-cell treatments. By casting dose finding as a Bayesian model selection problem, we propose an adaptive design by simultaneously incorporating the toxicity and efficacy outcomes to select the optimal biological dose (OBD) in phase I/II clinical trials. Without imposing any parametric assumption or shape constraint on the underlying dose response curves, we specify curve-free models for both the toxicity and efficacy endpoints to determine the OBD. By integrating the observed data across all dose levels, the proposed design is coherent in dose assignment and thus greatly enhances efficiency and accuracy in pinning down the right dose. Not only does our design possess a completely new yet flexible dose-finding framework, but it also has satisfactory and robust performance as demonstrated by extensive simulation studies. In addition, we show that our design enjoys desirable coherence prop-

erties, while most of existing phase I/II designs do not. We further extend the design to accommodate late-onset outcomes which are common in immunotherapy. The proposed design is exemplified with a phase I/II clinical trial in chronic lymphocytic leukemia.

A Simulation Study Evaluating Phase I Clinical Trial Designs for Combinational Agents

♦ *Shu Wang, Elias Sayour and Ji-Hyun Lee*

University of Florida
swang0221@ufl.edu

Combinational therapies that combine two or more therapeutic agents have become very common in cancer treatments. Dose-finding for combinational agents is considerably more complex than single agent, because only partial order of dose toxicity is known. In response, novel phase I clinical trial designs for combinational agents have been extensively proposed. However, with so many available designs, studies that compare their performances and explore the impact of design parameters, along with providing recommendations are limited. We evaluated available phase I designs that identify a single MTD for combinational agents using simulation studies under various scenarios. We also explored the influences of different design parameters and summarized the risks/benefits of each design to provide general guidance in design selection.

Bayesian Response Adaptive Randomization Design with A Composite Endpoint of Mortality and Morbidity

♦ *Zhongying Xu and Chung-Chou Chang*

University of Pittsburgh
zhx17@pitt.edu

If treatment allocation of patients during a trial is based on the observed responses and can be adapted sequentially, it could minimize the expected number of failures and maximize patients' benefits. In response-adaptive randomization (RAR), future treatment allocation ratios are determined based on the past treatment assignments and the response of patients collected before the decision time point. In this study, we developed a Bayesian RAR design targeting the endpoint of organ support-free days (OSFD) for patients admitted to the intensive care units (ICU). The OSFD is a mixture of mortality and morbidity assessed by the number of days free of organ support. In the past, researchers treated OSFD as an ordinal outcome variable that an arbitrary low number, for example, -1 or -100, is assigned to those who died in the ICU. We propose a novel RAR design for a composite endpoint of mortality and morbidity, e.g., OSFD, by using a Bayesian mixture model with a Markov chain Monte Carlo sampling to estimate the posterior probability of OSFD and determine treatment allocation ratios at each interim. Simulation was conducted to compare the performance of our proposed design under various randomization rules and different alpha spending functions. The results show that our RAR design using Bayesian inference benefits more patients while assuring adequate power for the target trial.

Session 6H : New Developments In Modern Nonparametric Statistics And The Applications

Doubly robust U-statistic with applications

♦ *Ao Yuan, Anqi Yin and Ming Tan*

Georgetown University
ay312@georgetown.edu

With the increased availability to capture large data, causal inference has received renewed attention and is playing an ever important

role in biomedicine and economics. However, one major methodological hurdle is that existing methods rely on many unverifiable assumptions. Thus robust modeling is a particularly important approach complementary to sensitivity analysis where different models are examined under various assumptions. The more robust an approach is with respect to model assumptions, the more meritorious it is. The doubly robust estimator (DRE) represents a major advance in this direction. However, in practice many outcome measures are functionals of multiple distributions, which can only be estimated via U-statistics and the existing DREs do not apply. In this article, we propose and study a broad class of semiparametric U-statistic DREs, which uses semiparametric specifications for the propensity score and outcome models in the construction of the U-statistic, to further enhance the robustness. Comprehensive asymptotic properties of the proposed estimators are investigated, extensive simulation studies are conducted to evaluate their finite sample behavior and compare with the corresponding parametric U-statistics and the naive estimators, which show significant advantages. Then the method is applied to analyzing a real data from the AIDS Clinical Trials Group.

Joint Semiparametric Models for Case-Cohort Designs

Weibin Zhong¹ and ♦Guoqing Diao²

¹Bristol Myers Squibb

²George Washington University
gdiao@email.gwu.edu

Two-phase studies such as case-cohort and nested case-control studies are widely used cost-effective sampling strategies. In the first phase, the observed failure/censoring time and inexpensive exposures are collected. In the second phase, a subgroup of subjects is selected for measurements of expensive exposures based on the information from the first phase. One challenging issue is how to utilize all the available information to conduct efficient regression analyses of the two-phase study data. This paper proposes a joint semiparametric modeling of the survival outcome and the expensive exposures. Specifically, we assume a class of semiparametric transformation models and a semiparametric density ratio model for the survival outcome and the expensive exposures, respectively. The class of semiparametric transformation models includes the proportional hazards model and the proportional odds model as special cases. The density ratio model is flexible in modeling multivariate mixed-type data. We develop efficient likelihood-based estimation and inference procedures and establish the large sample properties of the nonparametric maximum likelihood estimators. Extensive numerical studies reveal that the proposed methods perform well in practical settings. The proposed methods also appear to be reasonably robust under various model mis-specifications. An application to the National Wilms Tumor Study is provided.

Novel empirical likelihood inference for the mean difference with right-censored data

Kangni Alemjrodo¹ and ♦Yichuan Zhao²

¹Purdue University

²Georgia State University
yichuan@gsu.edu

This paper focuses on comparing two means and finding a confidence interval for the difference of two means with right-censored data using the empirical likelihood method combined with the independent and identically distributed random functions representation. In the literature, some early researchers proposed empirical link-based confidence intervals for the mean difference based on right-censored data using the synthetic data approach. However,

their empirical log-likelihood ratio statistic has a scaled chi-squared distribution. To avoid the estimation of the scale parameter in constructing confidence intervals, we propose an empirical likelihood method based on the independent and identically distributed representation of Kaplan-Meier weights involved in the empirical likelihood ratio. We obtain the standard chi-squared distribution. We also apply the adjusted empirical likelihood to improve coverage accuracy for small samples. In addition, we investigate a new empirical likelihood method, the mean empirical likelihood, within the framework of our study. The performances of all the empirical likelihood methods are compared via extensive simulations. The proposed empirical likelihood-based confidence interval has better coverage accuracy than those from existing methods. Finally, our findings are illustrated with a real data set.

Asymptotic Normality of Gini Correlation in High Dimension with Applications to the K-sample Problem

◆ *Yongli Sang¹ and Xin Dang²*

¹University of Louisiana at Lafayette

²University of Mississippi

yongli.sang@louisiana.edu

The categorical Gini correlation proposed by Dang et al. is a dependence measure between a categorical and a numerical variables, which can characterize independence of the two variables. The asymptotic distributions of the sample correlation under the dependence and independence have been established when the dimension of the numerical variable is fixed. However, its asymptotic distribution for high dimensional data has not been explored. In this paper, we develop the central limit theorem for the Gini correlation for the more realistic setting where the dimensionality of the numerical variable is diverging. We then construct a powerful and consistent test for the K-sample problem based on the asymptotic normality. The proposed test not only avoids computation burden but also gains power over the permutation procedure. Simulation studies and real data illustrations show that the proposed test is more competitive to existing methods across a broad range of realistic situations, especially in unbalanced cases.

Session 7A : Novel Statistical Methods For -Omic Data Analysis

LongStrain: An integrated strain-level analytic pipeline utilizing longitudinal metagenomics data

Boyan Zhou and ◆ *Huilin Li*

New York University

Huilin.Li@nyulangone.org

Despite the steady growth of longitudinal microbiome studies, most existing methods for strain-level analysis do not allow for the simultaneous interrogation of strain proportions and genome-wide variants in longitudinal metagenomic samples. We introduce LongStrain, an integrated pipeline for the analysis of metagenomic data from individuals with longitudinal or repeated samples. Our algorithm improves the efficiency and accuracy of strain identification by jointly modeling the strain proportion and genomic variants in combined multiple samples within individuals. With extensive simulation and real data analysis, we illustrate the superiority and potential applications of LongStrain.

An all-in-one statistical framework that simulates realistic single-cell omics data and infers cell heterogeneity structure

Jingyi Jessica Li

UCLA

lijy03@g.ucla.edu

The generation of realistic synthetic data is essential for benchmarking numerous computation tools developed for single-cell omics data. Here we propose an all-in-one statistical framework that generates single-cell omics data at both the read and count levels from various cell heterogeneity structures, including discrete cell types, continuous cell trajectories, and spatial cell locations. Our framework uses a unified probabilistic model with accessible likelihood. This probabilistic formulation is advantageous in that it enables a straightforward discernment of the heterogeneity structure that best fits a single-cell omics dataset, by leveraging the statistical model selection principle. Moreover, the ability to generate sequence reads, in addition to read counts, allows the benchmarking of low-level computational tools.

Data-Type Weighted Multi-Omics Spectral Clustering for Disease Subtyping

◆ *Peifeng Ruan and Hongyu Zhao*

Yale University

peifeng.ruan@yale.edu

Many multi-omics integration methods for disease subtyping have been proposed, with an intuition that incorporating more types of omics data produces better results. However, there are situations when integrating more omics data may negatively impact the performance of integration methods, such as including data type that is noisy in disease subtyping and redundant data types that may dominate the disease subtyping signals and decrease the subtyping accuracy. We propose a novel multi-omics regularized spectral clustering framework to integrate different omics data types, which learns the weights of each data type's signal for disease subtyping as well as the signal's redundancy level. Simulation studies and applications to multi-omics data of several cancer types from The Cancer Genome Atlas project suggest that the proposed multi-omics regularized spectral clustering framework achieves higher clustering accuracies and identifies new cancer subtypes that more accurately predict patient survival and are more biologically meaningful.

Deep Learning Methods for Retinal Imaging Genetics

Wei Chen

University of Pittsburgh

wec47@pitt.edu

Age-related Macular Degeneration (AMD) is a multifactorial irreversible retina disease and the leading cause of blindness in the developed world. Multiple factors including aging, genetics, and smoking are associated with AMD development and its progression. Successful genome-wide association studies (GWAS) of AMD have identified over 30 genes that are significantly associated with advanced AMD including dry and wet subtypes. Supported by the National Eye Institute and other resources, my group has assembled several large-scale image and genetics datasets including tens of thousands of individuals with hundreds of thousands of color fundus images. The combination of wealthy genetics and fundus image data, plus the well-characterized clinical phenotypes provides unprecedented opportunities to explore novel directions for studying retinal disease. In this talk, I will present our recent computational work to address several key issues and challenges in the analysis of such multi-modal data. I will discuss several models to predict AMD risk and progression using genetics data, image data, or both. We show that statistical and deep learning approaches are critical in understanding AMD pathogenesis and predicting disease progression. Our methods and findings will have potential to enhance the early prevention and current clinical management of the disease and

provide insights for novel precision treatment development.

Session 7B : Modern Time Series And Network Methods In Data Science.

Collaborative Spectral Clustering in Attributed Networks

Pengsheng Ji

Univ. of Georgia
psji@uga.edu

We proposed a novel spectral clustering algorithm for attributed networks, where n nodes split into R non-overlapping communities and each node has a p -dimensional meta covariate from various of formats such as text, image, speech etc. The connectivity matrix $W_{n \times n}$ is constructed with the adjacent matrix $A_{n \times n}$ and covariate matrix $X_{n \times p}$, and $W = (1 - \alpha)A + \alpha K(X, X')$, where $\alpha \in [0, 1]$ is a tuning parameter and K is a Kernel to measure the covariate similarities. We then perform the classical k -means algorithm on the element-wise ratio matrix of the first K leading eigenvector of W . Theoretical and simulation studies showed the consistent performance under both Stochastic Block Model (SBM) and Degree-Corrected Block Model (DCBM), especially in imbalanced networks where most community detection algorithms fail.

High Quantile Regression for Tail Dependent Time Series

Ting Zhang

University of Georgia
tingzhang@uga.edu

Quantile regression serves as a popular and powerful approach for studying the effect of regressors on quantiles of a response distribution. However, existing results on quantile regression were mainly developed when the quantile level is fixed, and the data are often assumed to be independent. Motivated by recent applications, we consider the situation where (i) the quantile level is not fixed and can grow with the sample size to capture the tail phenomena; and (ii) the data are no longer independent but collected as a time series that can exhibit serial dependence in both tail and non-tail regions. To study the asymptotic theory for high quantile regression estimators in the time series setting, we introduce a previously undescribed tail adversarial stability condition, and show that it leads to an interpretable and convenient framework for obtaining limit theorems for time series that exhibit serial dependence in the tail region but are not necessarily strong mixing. Numerical experiments are provided to illustrate the effect of tail dependence on high quantile regression estimators, where simply ignoring the tail dependence may lead to misleading p-values.

Dimension Reduction in Time Series Under the Presence of Conditional Heteroscedasticity

Murilo Dasilva, [♦]T. N. Sriram and Yuan Ke

University of Georgia
tn@uga.edu

We consider a time series $\{x_t; t \geq 1\}$, where the conditional mean of x_t is assumed to be an unknown function of linear combinations of past p observations, and the conditional variance of x_t is assumed to be an unknown function of linear combinations of past q squared residuals. The linear combinations are assumed to contain all the necessary information about x_t that is available through the conditional mean and conditional variance, respectively. Using Nadaraya-Watson kernel smoother to estimate the unknown mean and variance function, we propose an iterative estimation approach to estimate the parameter matrices associated with the linear combinations. The estimators are shown to be consistent. To overcome

computational challenges and provide numerical stability, we propose a new angular representation of parameter matrices. We examine the performance of the estimators through simulation studies and apply the iterative estimation procedure to model and forecast the Brazilian Real (BRL)/ U.S. Dollar Exchange Rate. For the BRL/USD series, we show that our estimated linear combinations yield a better time series model than an AR-ARCH model in terms of out-of-sample forecasts.

Multiple autocovariance changepoints problems in high-dimensional time series

Yuan Ke

University of Georgia
yuan.ke@uga.edu

We consider two problems about changepoints in autocovariance structures of a high-dimensional time series with heavy-tailed innovations. First, we study a multiple changepoints detection method based on the matrix max-norm of the tail-robust moving sum statistic. From a nonasymptotic perspective, we characterize the interplay among the window size, the dimensionality of moving sum statistic, and the minimal spacing, the value of the smallest change for underlying changepoints. While from an asymptotic perspective, under mild conditions, we show that the number and the locations of changepoints can be consistently estimated with a new data-driven threshold choice. Second, based on the same statistic and its null distribution constructed by block-wise permutations, we study the testing of changepoint at a prespecified location.

Session 7C : Innovative Approach Of Hidden Markov Model

Bayesian Semiparametric Hidden Markov Tensor Partition Models for Longitudinal Data with Local Variable Selection

Giorgio Paulon, Peter Mueller and [♦]Abhra Sarkar

UT-Austin
abhra.sarkar@utexas.edu

We present a flexible Bayesian semiparametric mixed model for longitudinal data analysis in the presence of potentially high-dimensional categorical covariates. Building on a novel hidden Markov tensor decomposition technique, our proposed method allows the fixed effects components to vary between dependent random partitions of the covariate space at different time points. The mechanism not only allows different sets of covariates to be included in the model at different time points but also allows the selected predictors' influences to vary flexibly over time. Smooth time-varying additive random effects are used to capture subject-specific heterogeneity. We establish posterior convergence guarantees for both function estimation and variable selection. We design a Markov chain Monte Carlo algorithm for posterior computation. We evaluate the method's empirical performances through synthetic experiments and demonstrate its practical utility through real-world applications.

Non-Standard Applications of Hidden Markov Models in the Biosciences

[♦]Jordan Aron¹, Matthew O. Gribble², Li C. Cheung³, and Paul Albert³

¹University of Minnesota

²University of Alabama at Birmingham School of Public Health

³National Cancer Center
Aron0064@umn.edu

Hidden Markov models (HMMs) have been used in a wide variety of applications in the biosciences including in genetics, modeling the natural history of disease, and in the environmental monitoring of toxins. Commonly, an E-M algorithm is used for maximum-likelihood estimation where a dynamic programming technique called the forward-backward algorithm is employed to make the E-step computations tractable. However, for some models the standard forward-backward algorithm cannot be applied, and novel adaptations are required for implementation. We begin by presenting the standard approach and then present two models that require novel dynamic programming methods for estimation. One example comes from disease modeling where the state-space is only partially observed (i.e., we only know that an individual is in one of a few possible states), while the second example is from environmental monitoring of toxins where the emissions distribution are not independent given the hidden state. We illustrate these approaches with actual data analysis from these examples.

A hidden Markov model approach for a joinpoint trend analysis

♦*Hyoyoung Choo-Wosoba, Philip Rosenburg and Paul Albert*

National Cancer Institute

hyoyoung.choo-wosoba@nih.gov

Joinpoint analyses has been widely used for the analysis of cancer incidence data to describe changes in patterns of disease incidence over time. The joinpoint model assumes piece-wise linear curves joined at multiple unknown joinpoints where the slope changes. However, this approach becomes computationally infeasible when the number of potential joinpoints and time points are large (e.g., joinpoints ≥ 4 , time points ≥ 50), since the estimation procedure involves enumerating all the possible joinpoints. We propose a hidden Markov model (HMM) that provides a computationally feasible approach for handling larger numbers of joinpoints/time points than is feasible with a standard approach. We show that it is natural to formulate the joinpoint model by modeling successive differences in observations as a HMM with the Markov chain being specified as a birth-process where each state change reflects the occurrence of a subsequent joinpoint. The standard Baum-Welch algorithm can be used for obtaining maximum-likelihood estimates as long as we assume that the estimated incidence is approximately normal with a constraint variance. In many realistic settings the variance in the estimated incidence increases with the mean. In this case, we develop a two-step algorithm for obtaining approximate maximum-likelihood estimators that updates the variances in the emissions distribution with successive applications of the Baum-Welch algorithm. We also generalize the approach for considering rates based on smaller disease counts by assuming a Poisson distribution for disease counts. We show the performance of the various estimation approaches with simulations and an example.

Session 7D : Statistical Advances And Applications In Analyzing Large Scale & Multi-Omic Single-Cell Data

iscTrack, a semi-supervised algorithm and interactive single-cell tool to track emerging transcriptional states in serial samples

Jiannong Li, Scott Cukras, Sathya Sriramareddy, Keiran Smalley, Xiaoping Yu and ♦Ann Chen

Moffitt Cancer Center

ann.chen@moffitt.org

Single-cell RNA sequencing (scRNA-seq) technique becomes available to characterize transcriptomic profiling and heterogeneity in

cancer research. The emerging evidence indicate that tumor transcriptional heterogeneity and plasticity plays an important role in drug response and resistance. We devised a multi-stage semi-supervised approach to analyze scRNA-seq data and track transcriptional state shifts, expansions, emerging and disappearing states in patient tumor samples over time. We created a responsive web tool, iscTrack, using Node.js as the API web server, Vue.js and the Quasar framework on the client side, and the D3 library for plots to interactively investigate the progression mechanism and provide treatment prediction using an enrichment approach. We conducted simulations to evaluate the performance of the proposed algorithm and evaluated its performance against commonly used unsupervised approach. We further applied the proposed method to analyze serial samples from melanoma patients and validate our treatment prediction using melanoma cell lines. The simulation results showed our approach was able to identify emerging state, specific to the progression sample with elevated biomarkers. For real patient datasets, our approach was able to select effective treatments to target resistant states, including a promising dual CDC7/CDK9 kinase inhibitor PHA-767491 and maternal and embryonic leucine zipper kinase MELK inhibitor OTSSP167, which were further validated experimentally using melanoma cell lines, SK-Mel28-RR and WM164. We showed that our proposed approach iscTrack is able to identify resistant states in patients' progressive samples and also provide promising therapies overcoming resistance.

Deep learning methods for cell type identification and gene expression imputation

Sijie Yao, Xiaoqing Yu and ♦Xuefeng Wang

Moffitt Cancer Center

xuefeng.wang@moffitt.org

The analysis of single-cell RNA sequencing (scRNA-seq) data faces challenges from cumbersome cell type identification processes and a large scale of gene expression dropout. In this talk, we first discuss a hallmark-biomarker-based automatic cell typing method based on wide and deep learning (WDL) framework. We show that DL can have superior classification accuracy when the training and testing of a model arise from the same cancer type but on different platforms. More specifically, WDL compared to traditional deep learning models can substantially increase the overall cell type prediction accuracy and T cell subtypes when the models were trained using melanoma data obtained from the 10X platform and tested on basal cell carcinoma data obtained using SMART-seq. In the second part of the talk, we will introduce a new gene expression imputation framework based on Generative Adversarial Imputation Nets (GAIN). Instead of estimating missing expression values of all genes, our method significantly improves computational efficiency by only imputing a subset gene panel that is most informative to distinguish cell types.

Nonparametric Interrogation of Transcriptional Regulation in Single-Cell RNA and Chromatin Accessibility Multiomic Data

Yuchao Jiang

UNC Chapel Hill

yuchaoj@email.unc.edu

Epigenetic control of gene expression is highly cell-type- and context-specific. Yet, despite its complexity, gene regulatory logic can be broken down into modular components consisting of a transcription factor (TF) activating or repressing the expression of a target gene through its binding to a cis-regulatory region. Recent advances in joint profiling of transcription and chromatin accessibility with single-cell resolution offer unprecedented opportunities

to interrogate such regulatory logic. Here, we propose a nonparametric approach, TRIPOD, to detect and characterize three-way relationships between a TF, its target gene, and the accessibility of the TF's binding site, using single-cell RNA and ATAC multiomic data. We apply TRIPOD to interrogate cell-type-specific regulatory logic in peripheral blood mononuclear cells and contrast our results to detections from enhancer databases, cis-eQTL studies, ChIP-seq experiments, and TF knockdown/knockout studies. We then apply TRIPOD to mouse embryonic brain data during neurogenesis and gliogenesis and identified known and novel putative regulatory relationships, validated by ChIP-seq and PLAC-seq. Finally, we demonstrate TRIPOD on SHARE-seq data of differentiating mouse hair follicle cells and identify lineage-specific regulation supported by histone marks for gene activation and super-enhancer annotations.

A statistical framework for scRNA-seq data modeling: simulation and applications

◆ *Guoshuai Cai*¹, *Xizhi Luo*¹, *Fei Qin*¹ and *Feifei Xiao*²

¹University of South Carolina

²University of Florida

caigs.whu@gmail.com

Recent advancements in single-cell RNA sequencing (scRNA-seq) have enabled time-efficient transcriptome profiling in individual cells. To optimize sequencing protocols and develop reliable analysis methods for various application scenarios, accurate statistical modeling scRNA-seq data is required. That is still a challenge due to the noisy nature of scRNA-seq data. Here, we introduce a new statistical framework for scRNA-seq analysis and show the improved application in data simulation and the inference of transcriptional bursting kinetics.

Session 7E : Modern Streaming Data Analysis: Process Monitoring

Fault Classification for High-dimensional Data Streams: A Directional Diagnostic Framework Based on Multiple Hypothesis Testing

Dongdong Xiang

East China Normal University

ddxiang@sfs.ecnu.edu.cn

In various modern statistical process control applications that involve high-dimensional data streams (HDDS), accurate fault diagnosis of out-of-control (OC) streams is becoming crucial. The existing diagnostic approaches either focus on moderate-dimensional processes or are unable to determine the shift direction accurately, especially when the signal-to-noise ratio is low. In this paper, we conduct a bold trial and consider the fault classification problem of the mean vector of HDDS where determining the shift direction of the OC streams is important to perform customized repairs. To this end, under the basic assumptions that the in-control data streams are normal with mean 0 and variance 1, and that the high-dimensional observations after the alarm are solely OC, the problem is formulated into a three-classification multiple testing framework, and an efficient data-driven diagnostic procedure is developed to minimize the expected number of false positives and to control the missed discovery rate at given level. The procedure is statistically optimal and computationally efficient, and improves the diagnostic effectiveness by considering directional information, which provides insights to guide further decisions. Both theoretical and numerical results reveal the superiority of the new method.

Adversarially Robust Sequential Hypothesis Testing

*Shuchen Cao*¹, ◆ *Ruizhi Zhang*¹ and *Shaofeng Zou*²

¹University of Nebraska-Lincoln

²University at Buffalo, The State University of New York

rzhang35@unl.edu

The problem of sequential hypothesis testing is studied, where samples are taken sequentially, and the goal is to distinguish between the null hypothesis where the samples are generated according to a distribution p and the alternative hypothesis where the samples are generated according to a distribution q . The defender (decision maker) aims to distinguish the two hypotheses using as fewer samples as possible subject to false alarm constraints. The problem is studied under the adversarial setting, where the data generating distributions under the two hypotheses are manipulated by an adversary, whose goal is to deteriorate the performance of the defender, e.g., increasing the probability of error and expected sample sizes, with minimal cost. This problem is formulated as a non-zero-sum game between the defender and the adversary. A pair of strategies (the adversary's strategy and the defender's strategy) is proposed and proved to be a Nash equilibrium pair for the non-zero-sum game between the adversary and the defender asymptotically. The defender's strategy is a sequential probability ratio test and thus is computationally efficient for practical implementation.

Recent advances in quality and industrial analytics

Fugee Tsung

HKUST

season@ust.hk

This talk will present and discuss the challenges and opportunities that quality and industrial analytics face in the era of digital transformation. There is an ample opportunity for quality and industrial analytics, under the digital transformation paradigm, to further explore ways of creating value from data and big data (e.g., data quality and safety assurance, big-data-driven process or product quality monitoring, improvement and optimization, fault diagnosis and risk management, consistent data fusion of several unstructured data sources, etc.). Emerging research issues such as change detection in heterogeneous data streams will be discussed.

Asymptotic Optimality Theory for Active Quickest Detection with Two Affected Streams

Qunzhi Xu

Georgia Institute of Technology

xuqunzhi@gatech.edu

Active quickest detection problems under the resource or sampling constraints have many real-world applications ranging from quality control in manufacturing to biosurveillance to security. Under a general setting, we monitor p local streams in a system under the sampling control constraint in the sense that we are only able to take observations from r of these p local streams at each time instant. Here we assume that at some unknown change time, an undesired event occurs to the system and changes the distributions of a subset of s unknown local streams. The objective is how to adaptively sample local streams and decide when to raise a global alarm, so that we can detect the correct change as quickly as possible subject to the false alarm constraint. In this paper, we develop the first asymptotic optimality theory in the active quickest detection literature for the case when $s = r = 2$. To be more concrete, we propose to combine three ideas to develop efficient active quickest detection algorithms: (1) win-stay, lose-switch sampling strategy; (2) local CUSUM statistics for local monitoring; and (3) the SUM-Shrinkage technique to fuse local statistics into a global decision. We show that our proposed algorithms are asymptotically optimal in the sense of mini-

mizing detection delay up to the second order subject to the false alarm constraint. Numerical studies are conducted to validate our theoretical results.

Session 7F : Discriminant And Cluster Analysis For Complex Data

Conditional probability tensor decompositions for multivariate categorical response regression

♦Aaron Molstad¹ and Xin Zhang²

¹University of Florida

²Florida State University
amolstad@ufl.edu

In many modern regression applications, the response consists of multiple categorical random variables whose probability mass is a function of a common set of predictors. In this article, we propose a new method for modeling such a probability mass function in settings where the number of response variables, the number of categories per response, and the dimension of the predictor are large. We introduce a latent variable model which implies a low-rank tensor decomposition of the conditional probability tensor. This model is based on the connection between the conditional independence of responses, or lack thereof, and the rank of their conditional probability tensor. Conveniently, our model can be interpreted in terms of a mixture of regressions and can thus be fit using maximum likelihood. We derive an efficient and scalable penalized expectation maximization algorithm to fit this model and examine its statistical properties. We demonstrate the encouraging performance of our method through both simulation studies and an application to modeling the functional classes of genes.

Quadratic Discriminant Analysis by Projection

Ruiyang Wu and ♦Ning Hao

University of Arizona
nhao@math.arizona.edu

Discriminant analysis, including linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), is a popular approach to classification problems. It is well known that LDA is suboptimal to analyze heteroscedastic data, for which QDA would be an ideal tool. However, QDA is less helpful when the number of features in a data set is moderate or high, and LDA and its variants often perform better due to their robustness against dimensionality. In this talk, we will introduce a new dimension reduction and classification method based on QDA. In particular, we define and estimate the optimal one-dimensional (1D) subspace for QDA, which is a novel hybrid approach to discriminant analysis. The new method can handle data heteroscedasticity with number of parameters equal to that of LDA. Therefore, it is more stable than the standard QDA and works well for data in moderate dimensions. We show an estimation consistency property of our method, and compare it with LDA, QDA, regularized discriminant analysis (RDA) and a few other competitors by simulated and real data examples.

A Doubly-Enhanced EM Algorithm for Model-Based Tensor Clustering

♦Qing Mai, Xin Zhang, Yuqing Pan and Kai Deng

Florida State University
qmai@fsu.edu

Modern scientific studies often collect data sets in the form of tensors. These datasets call for innovative statistical analysis methods. In particular, there is a pressing need for tensor clustering methods to understand the heterogeneity in the data. We propose a tensor

normal mixture model approach to enable probabilistic interpretation and computational tractability. Our statistical model leverages the tensor covariance structure to reduce the number of parameters for parsimonious modeling, and at the same time explicitly exploits the correlations for better variable selection and clustering. We propose a doubly-enhanced expectation-maximization (DEEM) algorithm to perform clustering under this model. Both the Expectation-step and the Maximization-step are carefully tailored for tensor data in order to maximize statistical accuracy and minimize computational costs in high dimensions. Theoretical studies confirm that DEEM achieves consistent clustering even when the dimension of each mode of the tensors grows at an exponential rate of the sample size. Numerical studies demonstrate favorable performance of DEEM in comparison to existing methods.

Stochastic Low-rank Tensor Bandits for Multi-dimensional Online Decision Making

Will Wei Sun

Purdue University
sun244@purdue.edu

Multi-dimensional online decision making plays a crucial role in many real applications such as online recommendation and digital marketing. In these problems, a decision at each time is a combination of choices from different types of entities. To solve it, we introduce stochastic low-rank tensor bandits, a class of bandits whose mean rewards can be represented as a low-rank tensor. We consider two settings, tensor bandits without context and tensor bandits with context. In the first setting, the platform aims to find the optimal decision with the highest expected reward, a.k.a, the largest entry of true reward tensor. In the second setting, some modes of the tensor are contexts and the rest modes are decisions, and the goal is to find the optimal decision given the contextual information. We propose two learning algorithms tensor elimination and tensor epoch-greedy for tensor bandits without context, and derive finite-time regret bounds for them. Comparing with existing competitive methods, tensor elimination has the best overall regret bound and tensor epoch-greedy has a sharper dependency on dimensions of the reward tensor. Furthermore, we develop a practically effective Bayesian algorithm called tensor ensemble sampling for tensor bandits with context. Numerical experiments back up our theoretical findings and show that our algorithms outperform various state-of-the-art approaches that ignore the tensor low-rank structure. In an online advertising application with contextual information, our tensor ensemble sampling reduces the cumulative regret by 75% compared to the benchmark method.

Session 7G : Design And Analysis In Vaccine Development And Its Challenges

Assessing the Role of Antibody in Vaccine Protection

Dean Follmann

NIH
dfollmann@niaid.nih.gov

COVID-19 vaccines have demonstrated outstanding efficacy against disease, and the role of vaccine induced antibody in protection is of substantial interest. A major role for antibody would support the use of antibody as a trigger for additional ‘booster’ doses and allow inferring efficacy in different populations, such as children. In this talk we provide a new approach to assess the amount of protection that is mediated by antibody. The key idea is to augment vaccine trial data with data from a COVID-19 prevention trial where the in-

intervention is passively infused antibodies, and thus any protective efficacy is due entirely to antibody. We decompose the total effect of vaccination into a part due to antibody with the reminder due to other vaccinal effects. Cox models with time-varying covariates are used to estimate how efficacy depends on decaying antibody and the proportion mediated derived. This approach differs from standard methods which only use vaccine trial data and require different assumptions.

Sensitivity Analysis for Evaluating Principal Surrogate Endpoints Relaxing the Equal Early Clinical Risk Assumption

♦ *Ying Huang, Yingying Zhuang and Peter Gilbert*

Fred Hutchinson Cancer Research Center
yhuang@fredhutch.org

We consider the evaluation of post-randomization immune response biomarkers as principal surrogate endpoints of a vaccine's protective effect, based on data from randomized vaccine trials. An important metric for quantifying a biomarker's principal surrogacy in vaccine research is the vaccine efficacy curve, which shows a vaccine's efficacy as a function of potential biomarker values if receiving vaccine, among an "early always at risk" principal stratum of trial participants who remain disease-free at the time of biomarker measurement whether having received vaccine or placebo. Earlier work in principal surrogate evaluation relied on an "equal-early-clinical-risk" assumption for identifiability of the vaccine curve, based on observed disease status at the time of biomarker measurement. This assumption is violated in the common setting that the vaccine has an early effect on infection before the biomarker is measured. In particular, a vaccine's early protective effect observed in two phase III dengue vaccine trials (CYD14/CYD15) has motivated our current research development. In this project we relax the "equal-early clinical-risk" assumption and propose a new sensitivity analysis framework for principal surrogate evaluation allowing for early vaccine efficacy. Under this framework, we develop inference procedures for vaccine efficacy curve estimators based on the estimated maximum likelihood approach. We use the proposed methodology to assess the surrogacy of post-randomization neutralization titer in dengue vaccine trials.

Durability of Covid-19 Vaccines

Yu Gu

University of North Carolina
yugu@live.unc.edu

Evaluating the durability of protection afforded by Covid-19 vaccines is a public health priority, with the results needed to inform policies around booster vaccinations as well as those around non-pharmaceutical interventions. In this talk, I will present a general framework for estimating the effects of Covid-19 vaccines over time in phase 3 clinical trials and observational studies. I will show some results on the duration of vaccine protection from the Moderna pivotal trial and from the North Carolina statewide surveillance data. The latter data, which were published in the *New England Journal of Medicine* in January, provided rich information about the effectiveness of the Pfizer, Moderna, and Johnson & Johnson vaccines in reducing the risks of Covid-19, hospitalization, and death over time. I will discuss the implications of these results for booster vaccinations.

Statistical Consideration for Accelerated COVID-19 Vaccine Clinical Development in the Pandemic

James Zhou

HHS/ASPR/BARDA
james.zhou@hhs.gov

The COVID-19 pandemic presents an extraordinary challenge to global health. From the beginning the US Government (USG) has been supporting and collaborating with commercial vaccine manufacturers developing COVID-19 vaccine candidates using different platform technologies including mRNA, recombinant protein, and viral vectored vaccines. A showcase would be the accelerated clinical development of Moderna mRNA-1273 COVID-19 vaccine in parallel to the manufacturing production at risk. Statistical considerations have been meticulously elaborated and closely monitored along with expedited clinical protocol development and execution. I will illustrate critical statistical design and monitoring features of the Moderna Coronavirus Efficacy (COVE) Phase 3 Trial, and its subsequent adaptations to cross-over vaccination in participants received placebo in the blinded phase of mRNA-1273 primary series and booster vaccination in all participants as the response to the pandemic evolves.

Session 7H : Methods For Inference On Variable Importance Using Machine Learning

Inference for model-agnostic variable importance

♦ *Brian Williamson¹, Susan Shortreed¹, Peter Gilbert², Noah Simon³ and Marco Carone³*

¹Kaiser Permanente Washington Health Research Institute

²Fred Hutchinson Cancer Research Center

³University of Washington

brian.d.williamson@kp.org

In many applications, it is of interest to assess the relative contribution of features (or subsets of features) toward the goal of predicting a response; in other words, to gauge the variable importance of features. Most recent work on variable importance assessment has focused on describing the importance of features within the confines of a given prediction algorithm. However, such an assessment does not necessarily characterize the prediction potential of features and may provide a misleading reflection of the intrinsic value of these features. To address this limitation, I propose a general framework for nonparametric inference on interpretable algorithm-agnostic variable importance. I also discuss several approaches to summarizing the longitudinal importance of variables. I will illustrate the use of the proposed framework in the context of a prevention efficacy trial of an antibody against HIV-1 infection.

Variable importance measure for spatial machine learning models with application to air pollution exposure prediction

♦ *Si Cheng, Ali Shojaie, Lianne Sheppard and Adam Szpiro*

University of Washington

chengsi@uw.edu

Adequate exposure assessment is fundamental to air pollution studies and environmental epidemiology in general. It is often of interest to not only accurately predict the exposure, but also understand the mechanism captured by the model fit to the data. However, the latter is not always straightforward due to the complex nature of machine learning methods as well as the lack of unifying notions of variable importance given the diversity of models. This is further complicated in spatial modeling by the presence of spatial correlation. We tackle these problems in a study of the United States national PM_{2.5} sub-species data and Seattle traffic-related air pollution data. We propose an intuitive leave-one-out approach for variable importance that leads to interpretable and comparable measures across different models. The proposed variable importance measure clearly highlights the difference in mechanisms captured by different models,

even for those producing similar predictions. Observations from such measure could lead to more thorough scientific understanding as well as more informed decision-making from spatial analyses.

Floodgate: inference for model-free variable importance

◆ *Lu Zhang and Lucas Janson*

Harvard University
lu.zhang@g.harvard.edu

Many modern applications seek to understand the relationship between an outcome variable Y and a covariate X in the presence of a (possibly high-dimensional) confounding variable Z . Although much attention has been paid to testing whether Y depends on X given Z , in this paper we seek to go beyond testing by inferring the strength of that dependence. We first define our estimand, the minimum mean squared error (mMSE) gap, which quantifies the conditional relationship between Y and X in a way that is deterministic, model-free, interpretable, and sensitive to nonlinearities and interactions. We then propose a new inferential approach called floodgate that can leverage any working regression function chosen by the user (allowing, e.g., it to be fitted by a state-of-the-art machine learning algorithm or be derived from qualitative domain knowledge) to construct asymptotic confidence bounds, and we apply it to the mMSE gap. In addition to proving floodgate's asymptotic validity, we rigorously quantify its accuracy (distance from confidence bound to estimand) and robustness. We then show we can apply the same floodgate principle to a different measure of variable importance when Y is binary. Finally, we demonstrate floodgate's performance in a series of simulations and apply it to data from the UK Biobank to infer the strengths of dependence of platelet count on various groups of genetic mutations.

Regularization on Ensembles of Tree and Variable importance

◆ *Siyu Zhou and Lucas Mentch*

University of Pittsburgh
siz25@pitt.edu

Despite their well-established record, a full and satisfying explanation for the success of random forests has yet to be put forth. Here, we take a step in this direction. Comparing against bagging with non-randomized base learners, we demonstrate that random forests are implicitly regularized by the additional randomness injected into individual trees, making them highly advantageous in low signal-to-noise (SNR) settings. Furthermore, we show that this regularization property is not unique to tree-based ensembles and can be generalized to other supervised learning procedures. Motivated by this, we find that another surprising and counterintuitive means of regularizing ensembles can come from the inclusion of additional random noise features in the model. Importantly, this leads to substantial concerns about common notations of variable importance based on improved model accuracy, as even purely random noise can routinely register as statistically significant.

Session 8A : Ultra-High Dimensional Variable Selection And Zero-Inflated Negative Binomial Spatial And Temporal Regression

Spatiotemporal Zero-Inflated Bayesian Negative Binomial Regression Using Nearest Neighbor Gaussian Process and Polya-Gamma Mixtures

◆ *Qing He and Hsin-Hsiung Huang*

University of Central Florida
carsonqing@knights.ucf.edu

Spatiotemporal data analysis with massive zeros are widely used in many areas such as epidemiology and public health. We use a Bayesian framework to fit zero-inflated negative binomial models which introduces a set of latent variables from Polya-Gamma distributions. The posterior Markov chain Monte Carlo algorithm is efficient through Gibbs sampling. The proposed model accommodates varying spatial and temporal random effects through Gaussian processes, which give a flexible covariance structure for the random effects but have computation challenges when the dataset is large. To conquer the computation bottleneck in the Gaussian process, we adopt the nearest-neighbor Gaussian process method which approximates the kernel matrix using local experts. For the simulation study, we adopt multiple settings with varying sizes of spatial locations to evaluate the performance of the proposed model such as spatial and temporal random effects estimation and compare the result to other methods. We also apply the proposed algorithm to the COVID-19 data to examine death rates among counties with high and low social vulnerability levels in Florida, USA.

An Exchangeable Prior on Partitions for Clustering

◆ *Charles Harrison, Qing He and Hsin-Hsiung Huang*

University of Central Florida
charleswharrison@knights.ucf.edu

We propose a prior distribution on partitions for clustering. The prior is based on pairwise distances between observations, eliminates the need for analysts to know or estimate the number of clusters, and it facilitates parallel computation via exchangeability. We incorporate the prior with a matrix normal likelihood function and apply the resulting Gibbs sampler to a variety of simulated and real-world datasets.

Multi-Omics Integrative Analysis for Incomplete Data Using Weighted p -value Adjustment Approaches

Wenda Zhang¹, Joshua Habiger², Hsin-Hsiung Huang³ and ◆Yen-Yi Ho¹

¹University of South Carolina

²Oklahoma State University

³University of Central Florida
hoyen@stat.sc.edu

The advancements in high-throughput technologies provide exciting opportunities to obtain multi-omics data from the same individuals in a biomedical study since joint analyses of data from multiple sources offer many benefits. However, the occurrence of missing values is an inevitable issue in multi-omics data because measurements such as mRNA gene expression levels often require invasive tissue sampling from patients. Common approaches for addressing missing measurements include analyses based on observations with complete data or multiple imputation methods. In this paper, we propose a novel integrative multi-omics analytical framework based on p -value weight adjustment in order to incorporate observations with incomplete data into the analysis. By splitting the data into a complete set with full information and an incomplete set with missing measurements, we introduce mechanisms to derive weights and weight-adjusted p -values from the two sets. Through simulation analyses, we demonstrate that the proposed framework achieves considerable statistical power gains compared to a complete case analysis or multiple imputation approaches. We illustrate the implementation of our proposed framework in a study of preterm infant birth weights by a joint analysis of DNA methylation, mRNA, and the phenotypic outcome.

Sparse Bayesian Matrix-variate Regression with High-

dimensional Binary Response Data♦*Hsin-Hsiung Huang*¹ and *Shao-Shuan Wang*²¹University of Central Florida²National Central University
hsin.huang@ucf.edu

We propose a Bayesian generalized linear models for matrix-valued covariates data with shrinkage priors for estimation and variable selection in high-dimensional settings where the dimensions of the covariates increase as the sample size increases. This study is motivated by extending Bayesian approaches in a classical multivariate linear model. The proposed estimation can be applied to classifying matrix data such as images. We show that the proposed model achieves strong posterior consistency when the dimension grows at a sub-exponential rate with the sample size. Furthermore, we quantify the posterior contraction rate at which the posterior shrinks around the true regression coefficients. Simulation studies and an application to Electroencephalography and Leucorrhea data show the superior performance of the proposed method over the existing approaches.

Sparse Bayesian Matrix-variate Regression with High-dimensional Binary Response Data♦*Shao-Hsuan Wang*¹ and *Hsin-Hsiung Huang*²¹National Central University²University of Central Florida
picowang@gmail.com

We propose a Bayesian generalized linear models for matrix-valued covariates data with shrinkage priors for estimation and variable selection in high-dimensional settings where the dimensions of the covariates increase as the sample size increases. This study is motivated by extending Bayesian approaches in a classical multivariate linear model. The proposed estimation can be applied to classifying matrix data such as images. We show that the proposed model achieves strong posterior consistency when the dimension grows at a subexponential rate with the sample size. Furthermore, we quantify the posterior contraction rate at which the posterior shrinks around the true regression coefficients. Simulation studies and an application to Electroencephalography and Leucorrhea data show the superior performance of the proposed method over the existing approaches.

Session 8B : Recent Developments In Functional Data Analysis**Multilevel Modeling of Spatially Nested Functional Data: Spatiotemporal Patterns of Hospitalization Rates in the U.S. Dialysis Population***Yihao Li*¹, *Danh Nguyen*², *Sudipto Banerjee*¹, *Connie Rhee*², *Kam-yar Kalantar-Zadeh*², *Esra Kurum*³ and ♦*Damla Senturk*¹UCLA²UCI³UCRiverside

dsenturk@g.ucla.edu

End-stage renal disease patients on dialysis experience frequent hospitalizations. In addition to known temporal patterns of hospitalizations over the life span on dialysis, where poor outcomes are typically exacerbated during the first year on dialysis, variations in hospitalizations among dialysis facilities across the U.S. contribute to spatial variation. Utilizing national data from the United States Renal Data System (USRDS), we propose a novel multilevel spatiotemporal functional model to study spatiotemporal patterns of

hospitalization rates among dialysis facilities. Hospitalization rates of dialysis facilities are considered as spatially nested functional data with longitudinal hospitalizations nested in dialysis facilities and dialysis facilities nested in geographic regions. A multilevel Karhunen-Loeve expansion is utilized to model the two-level (facility and region) functional data, where spatial correlations are induced among region-specific principal component scores accounting for regional variation. A new efficient algorithm based on functional principal component analysis and Markov Chain Monte Carlo is proposed for estimation and inference. We report a novel application using USRDS data to characterize spatiotemporal patterns of hospitalization rates for over 400 health service areas across the U.S. and over the post-transition time on dialysis. Finite sample performance of the proposed method is studied through simulations.

Online Estimation for Functional Data*Fang Yao*

Beijing University

fyao@math.pku.edu.cn

Functional data analysis has attracted considerable interest, and is facing new challenges of the increasingly available data in streaming manner. In this work, we propose a new online method to dynamically update the local linear estimates of mean and covariance functions of functional data, which is the foundation of subsequent analysis. The kernel-type estimates can be decomposed into two sufficient statistics depending on the data-driven bandwidths. We propose to approximate the future optimal bandwidths by a dynamic sequence of candidates and combine the corresponding statistics across blocks to make an updated estimation. The proposed online method is easy to compute based on the stored sufficient statistics and current data block. Based on the asymptotic normality of the online mean and covariance function estimates, the relative efficiency in terms of integrated mean squared error is studied and a theoretical lower bound is obtained. This bound provides insight into the relationship between estimation accuracy and computational cost driven by the length of candidate bandwidth sequence that is pivotal in the online algorithm. Simulations and real data applications are provided to support such findings and show the advantages of the proposed method.

Functional ANOVA for High-Dimensional Spectral Analysis♦*Robert Krafty*¹, *Marie Tuft*², *Fabio Ferrarelli*³, *Ori Rosen*⁴ and *Zeda Li*⁵¹Emory University²Sandia National Laboratory³University of Pittsburgh⁴University of Texas⁵Baruch College, City University of New York
rkrafty@emory.edu

Technological advances have led to an increase in the breadth and number of studies that collect high-dimensional time series signals, such as EEG, from multiple groups and whose scientific goal is to understand differences in time series spectra between the groups. Although methods have been proposed for comparing populations of power spectra that are univariate functions of frequency, often referred to as analysis of power (ANOPOW), none exist when time series are high-dimensional and spectra are complex Hermitian matrix-valued functions. In this talk, we discuss a non-parametric Bayesian approach for ANOPOW with high-dimensional time series. The method models the collection of time series through a novel functional mixed effects factor model that can capture spectral differences between groups while accounting for within-group

spectral variability. The approach is motivated by and used to analyze resting-state high-dimensional EEG in patients hospitalized for a first psychotic episode to understand how their electrophysiology differs from that of healthy controls.

Factor-augmented model for functional data

Yuan Gao¹, Han Lin Shang² and ♦Yanrong Yang¹

¹The Australian National University

²Macquarie University
yanrong.yang@anu.edu.au

We propose modeling raw functional data as a mixture of a smooth function and a high-dimensional factor component. The conventional approach to retrieving the smooth function from the raw data is through various smoothing techniques. However, the smoothing model is inadequate to recover the smooth curve or capture the data variation in some situations. These include cases where there is a large amount of measurement error, the smoothing basis functions are incorrectly identified, or the step jumps in the functional mean levels are neglected. A factor-augmented smoothing model is proposed to address these challenges, and an iterative numerical estimation approach is implemented in practice. Including the factor model component in the proposed method solves the aforementioned problems since a few common factors often drive the variation that cannot be captured by the smoothing model. Asymptotic theorems are also established to demonstrate the effects of including factor structures on the smoothing results. Specifically, we show that the smoothing coefficients projected on the complement space of the factor loading matrix are asymptotically normal. As a byproduct of independent interest, an estimator for the population covariance matrix of the raw data is presented based on the proposed model. Extensive simulation studies illustrate that these factor adjustments are essential in improving estimation accuracy and avoiding the curse of dimensionality. The superiority of our model is also shown in modeling Australian temperature data.

Session 8C : Recent Advances In Robust Statistical Models For Censored And Missing Data

Censored autoregressive regression models with Student-t innovations

Katherine Andreina Loovaleriano¹, ♦Fernanda Langschumacher², Christian Galarza³ and Larissa Avilamatos¹

¹University of Campinas

²Ohio State University

³Escuela Superior Politécnica del Litoral
fernandalschumacher@gmail.com

Data collected over time is common in applications and may contain censored or missing observations, making it impossible to use standard statistical procedures. This work proposes an algorithm to estimate the parameters of a censored linear regression model when the regression errors are serially correlated and the innovations follow a Student-t distribution, which is widely used in statistical modeling of data containing outliers since its longer-than-normal tails provide a robust approach to handling such data. To fit the proposed model, maximum likelihood estimates are obtained through a stochastic approximation of the EM algorithm. The methods are applied to an environmental dataset regarding phosphorus concentration, which is subjected to a limit of detection and contains missing observations due to a program suspension caused by a temporary lack of funding.

Lasso regularization for censored skew-t regression and high-dimensional predictors

Victor Hugo Lachos

University of Connecticut
hlachos@uconn.edu

The skew-t distribution is an attractive family of asymmetrical heavy-tailed densities that includes the normal, skew-normal and Student's-t distributions as special cases. In this work, we propose a EM algorithm to handle skew-t censored regression models with a Lasso regularization to accommodate high dimensional predictors. We specify how this technique can be easily implemented using available R packages. The proposed methods are assessed using simulated and real data.

Extending multivariate Student's-t semiparametric mixed models for longitudinal data with censored responses and heavy tails

Thalita Mattos¹, Victor Hugo Lachos², Luis Mauricio Castro³ and ♦Larissa Matos¹

¹Universidade Estadual de Campinas

²University of Connecticut

³Pontificia Universidad Católica de Chile
larissam@unicamp.br

This paper extends the semi-parametric mixed model for longitudinal censored data with Gaussian errors by considering the Student's t-distribution. This model allows us to consider a flexible, functional dependence of an outcome variable over the covariates using non-parametric regression. Moreover, the proposed model takes into account the correlation between observations by using random effects. Penalized likelihood equations are applied to derive the maximum likelihood estimates that appear to be robust against outlying observations with respect to the Mahalanobis distance. We estimate non-parametric functions using smoothing splines under an EM-type algorithm framework. Finally, the proposed approach's performance is evaluated through extensive simulation studies and an application to two data sets from acquired immunodeficiency syndrome (AIDS) clinical trials.

Session 8D : Recent Advances In Latent Variable Analysis

A Note on Statistical Inference for Noisy Incomplete Binary Matrix

Gongjun Xu

University of Michigan
gongjun@umich.edu

We consider the statistical inference for noisy incomplete binary matrix completion, most of the categorical matrix completion literature focus on point estimation and prediction. This work moves one step further towards the statistical inference for binary matrix completion. Under a popular nonlinear unidimensional factor analysis model, we obtain a point estimator and derive its asymptotic distribution for any linear form of the probability matrix and latent factor scores. Moreover, our analysis adopts a flexible missing-entry design that does not require a random sampling scheme as required by most of the existing asymptotic results for matrix completion. The proposed estimator is statistically efficient and optimal, in the sense that the Cramer-Rao lower bound is achieved asymptotically for the model parameters. An application in linking two forms of an educational test is used to illustrate the theoretical results. This is a joint work with Yunxiao Chen and Chengcheng Li.

VEMIRT: A Variational EM Algorithm-based Shiny App for High-dimensional IRT Applications

♦ *Chun Wang*¹, *Gongjun Xu*², *Chenchen Ma*², *Ruoyi Zhu*¹ and *Jiaying Xiao*¹

¹University of Washington

²University of Michigan
wang4066@uw.edu

The increasing availability of rich survey data and the emerging needs of assessing multifaceted constructs in social science pose great challenges to existing techniques used to handle and analyze the large-scale assessment (LSA) data. We created a new R Shiny App called VEMIRT, which uses innovative Gaussian variational expectation-maximization (GVEM) methods to efficiently calibrate high-dimensional IRT models. In this talk, we will demonstrate a few snippets of GVEM in a variety of application scenarios, including (1) using an important-sampling enhanced GVEM algorithm to improve parameter recovery precision in confirmatory MIRT; (2) using a multiple-group GVEM with regularization to detect differential item functioning, and (3) using a bootstrap corrected GVEM on two commonly seen LSA designs, the balanced incomplete block designs, and multistage testing designs.

A random effect hidden Markov model for process data

Xueying Tang

University of Arizona
xytang@math.arizona.edu

Response process data from computer-based problem-solving items have received significant attention recently. Such data describe a respondent's problem-solving process as a sequence of timestamped actions. They are rich in information but are complex in structure. Several feature extraction methods have been developed for process data to overcome the difficulty brought by the non-standard data format. Although the extracted features preserve the information in response processes and are helpful for improving the accuracy of assessments, it is not easy to identify how latent traits are related to the actions. In this talk, we present a hidden Markov model (HMM) of the dynamics of problem-solving processes. The hidden states in HMM mimic the problem-solving stages. Respondents' latent traits are introduced in the model as random effects and are directly connected to the choice of actions, leading to straightforward interpretations of model parameters. We demonstrate how the proposed model can be used to understand problem-solving processes and measure respondents' latent traits through a case study of PISA process data.

Tree-informed Bayesian multi-source domain adaptation

♦ *Zhenke Wu*¹, *Zehang Li*², *Irena Chen*¹ and *Mengbing Li*¹

¹University of Michigan, Ann Arbor

²University of California, Santa Cruz
zhenkewu@umich.edu

Determining causes of deaths (COD) occurred outside of civil registration and vital statistics systems is challenging. A technique called verbal autopsy (VA) is widely adopted to gather information on deaths in practice. A VA consists of interviewing relatives of a deceased person about symptoms of the deceased in the period leading to the death, often resulting in multivariate binary responses. While statistical methods have been devised for estimating the cause-specific mortality fractions (CSMFs) for a study population, continued expansion of VA to new populations (or "domains") necessitates approaches that recognize between-domain differences while capitalizing on potential similarities. In this paper, we propose such a domain-adaptive method that integrates external between-domain similarity information encoded by a pre-specified rooted weighted

tree. Given a cause, we use latent class models to characterize the conditional distributions of the responses that may vary by domain. We specify a logistic stick-breaking Gaussian diffusion process prior along the tree for class mixing weights with node-specific spike-and-slab priors to pool information between the domains in a data-driven way. Posterior inference is conducted via a scalable variational Bayes algorithm. Simulation studies show that the domain adaptation enabled by the proposed method improves CSMF estimation and individual COD assignment. We also illustrate and evaluate the method using a validation data set. The paper concludes with a discussion on limitations and future directions.

Session 8E : New Advances In Microbiome Related Data Analysis

LinDA: Linear Models for Differential Abundance Analysis of Microbiome Compositional Data

♦ *Jun Chen*¹ and *Xianyang Zhang*²

¹Mayo Clinic

²Texas A&M University
Chen.Jun2@mayo.edu

Differential abundance analysis, which aims to identify microbial taxa whose abundance covaries with a variable of interest, is at the center of statistical analyses of microbiome data. Although the main interest is in drawing inferences on the absolute abundance, i.e., the number of microbial cells per unit area/volume at the ecological site such as the human gut, the data from a sequencing experiment reflects only the taxa relative abundance in a sample. Thus, microbiome data are compositional in nature. Analysis of such compositional data is challenging since the change in the absolute abundance of one taxon will lead to changes in the relative abundances of other taxa, making false positive control difficult. Here we present a simple, yet robust and highly scalable approach to tackle the compositional effects in differential abundance analysis. The method only requires the application of established statistical tools. It fits linear regression models on the centered log-ratio transformed data, identifies a bias term due to the transformation and compositional effect, and corrects the bias using the mode of the regression coefficients. Due to the algorithmic simplicity, our method is 100-1000 times faster than the state-of-the-art method ANCOM-BC. Under mild assumptions, we prove its asymptotic FDR control property, making it the first differential abundance method that enjoys a theoretical FDR control guarantee. The proposed method is very flexible and can be extended to mixed-effect models for the analysis of correlated microbiome data. Using comprehensive simulations and real data applications, we demonstrate that our method has overall the best performance in terms of FDR control and power among the competitors.

Identifying Microbial Interaction Networks Based on Irregularly Spaced Longitudinal Data

♦ *Jiang Gui*¹, *Jie Zhou*¹, *Weston Viles*² and *Annie Hoen*¹

¹Dartmouth College

²University of Southern Maine
Jiang.Gui@dartmouth.edu

We consider the selection of high-dimensional Gaussian graphical model when the data are irregularly spaced longitudinal measurements. Correlation functions are employed to account for the possible correlation among the measurements for the given subject. Two scenarios are considered, in which the first one assumes homogeneous correlation function among subjects, while the second one

allows heterogeneous correlation function. In both scenarios, we propose iterative algorithms to select the Gaussian graphical model from L1-penalized likelihood function. The algorithms take advantage of the popular graphical lasso algorithm and can be readily implemented. Simulation studies show that our algorithms can outperform the popular conventional algorithms such as graphical lasso when the correlation between measurements are not negligible. A real microbiome abundance dataset also is analyzed using our algorithm.

Synergy Regression of Microbiome and Metabolome Data

Yue Wang

Arizona State University
ywan1282@asu.edu

Mechanistic understanding of the microbiome requires identifying co-regulating microbial markers (e.g. taxa, metabolites, etc.) that are associated with host health outcomes. In this talk, I will discuss a new systems biology approach for regression analysis of multi-view microbiome data (e.g. metagenomics, metabolomics, etc.). Our method identifies multivariate association between the outcome and the latent factors common to all omics layers, and reveals specific variables underlying the multivariate association. I will illustrate the merit of the proposed method using an analysis of metagenomic and metabolomic data from the Study of Latinos.

A Novel Causal Mediation Analysis Approach for Zero-Inflated Count Mediators

♦Meilin Jiang¹, Seonjoo Lee², A. James O'malley³, Yaakov Stern² and Zhigang Li¹

¹University of Florida

²Columbia University

³Geisel School of Medicine at Dartmouth
meilin.jiang@ufl.edu

Mediation analyses play important roles in making causal inference in biomedical research to examine causal pathways that may be mediated by one or more intermediate variables (i.e., mediators). Although mediation frameworks have been well established such as counterfactual-outcomes (i.e., potential-outcomes) models and traditional linear mediation models, little effort has been devoted to dealing with mediators with zero-inflated structures due to challenges associated with excessive zeros. We develop a novel mediation modeling approach to address zero-inflated count mediators containing true zeros and (possibly) false zeros. The new approach can decompose the total mediation effect into two components induced by zero-inflated structures: the first component is attributable to the change in the mediator on its numerical scale and the second component is attributable to its binary change from zero to a non-zero status. An extensive simulation study is conducted to assess the performance and it shows that the proposed approach outperforms existing standard causal mediation analysis approaches. We also showcase the application of the proposed approach to a real study in comparison with a standard causal mediation analysis approach.

Session 8F : Statistical Computation Of Big Data With Biomedical Applications

Bayesian data compression

♦Rajarshi Guhaniyogi¹ and Aaron Scheffler²

¹Texas A & M University

²UC San Francisco
rajguhaniyogi@tamu.edu

Bayesian computation of high dimensional linear regression models with a popular Gaussian scale mixture prior distribution using Markov Chain Monte Carlo (MCMC) or its variants can be extremely slow or completely prohibitive due to the heavy computational cost that grows in the cubic order of p , with p as the number of features. Although a few recently developed algorithms make the computation efficient in presence of a small to moderately large sample size (with the complexity growing in the cubic order of n), the computation becomes intractable when sample size n is also large. In this talk we discuss an approach to compress the n original samples by a random linear transformation to m dimensions, and compute Bayesian regression with Gaussian scale mixture prior distributions with the randomly compressed response vector and feature matrix. Our proposed approach yields computational complexity growing in the cubic order of m . Another important motivation for this compression procedure is that it anonymizes the data by revealing little information about the original data in the course of analysis. We will show empirical investigation with the Horseshoe prior from the class of Gaussian scale mixture priors demonstrating closely similar inference and a massive reduction in per iteration computation time of the proposed approach compared to the regression with the full sample. We will also discuss posterior contraction rate for high dimensional predictor coefficient with a general class of shrinkage priors on them under data compression.

A 'Divide-and-Conquer' AECM Algorithm for Large non-Gaussian Longitudinal Data

♦Reuben Retnam¹, Sanvesh Srivastava² and Dipankar Bandyopadhyay¹

¹Virginia Commonwealth University

²University of Iowa
rpr2151@gmail.com

Features of non-Gaussianity, manifested via skewness and heavy tails, are ubiquitous in databases generated from large scale observational studies. Yet they continue to be routinely analyzed via linear/non-linear mixed effects models under standard Gaussian assumptions for the random terms. In periodontal disease data, these issues are applicable to the modeling of clinical attachment level and pocket depth. These problems are maintained, if not exacerbated, in the longitudinal data framework. In this research, we define and elucidate an extension of the skew-t linear mixed model suitable for a big data setting. This extensibility is achieved via the implementation of divide-and-conquer techniques that utilize the distributed expectation-maximization algorithm. Specifically, the E-steps of the AECM algorithm are run in parallel on multiple worker processes, while manager processes perform the M-steps with an updated fraction of the results from the local expectation steps. We prove convergence properties of this algorithm and show examples of its performance compared to traditional modelling methods on real and simulated data.

Bayesian Generalized Sparse Symmetric Tensor-on-Vector Regression

♦Sharmistha Guha and Rajarshi Guhaniyogi

Texas A&M University
sharmistha@tamu.edu

Motivated by brain connectome datasets acquired using diffusion weighted magnetic resonance imaging (dMRI), this article proposes a novel generalized Bayesian linear modeling framework with a symmetric tensor response and scalar predictors. The symmetric tensor coefficients corresponding to the scalar predictors are embedded with two features: low-rankness and group sparsity within

the low-rank structure. Besides offering computational efficiency and parsimony, these two features enable identification of important "tensor nodes" and "tensor cells" significantly associated with the predictors, with characterization of uncertainty. The proposed framework is empirically investigated under various simulation settings and with a real brain connectome dataset.

Ordinal Causal Discovery

♦*Yang Ni and Bani Mallick*

Texas A&M University
yni@stat.tamu.edu

Causal discovery for purely observational, categorical data is a long-standing challenging problem. Unlike continuous data, the vast majority of existing methods for categorical data focus on inferring the Markov equivalence class only, which leaves the direction of some causal relationships undetermined. This paper proposes an identifiable ordinal causal discovery method that exploits the ordinal information contained in many real-world applications to uniquely identify the causal structure. The proposed method is applicable beyond ordinal data via data discretization. Through real-world and synthetic experiments, we demonstrate that the proposed ordinal causal discovery method combined with simple score-and-search algorithms has favorable and robust performance compared to state-of-the-art alternative methods in both ordinal categorical and non-categorical data. An accompanied R package OCD is freely available at https://web.stat.tamu.edu/yni/files/OCD_0.1.0.tar.gz.

Session 8G : Recent Development In Innovative Clinical Trial Designs

Graphical representation of the Hochberg procedure and other equally weighted tests

♦*Dong Xi¹ and Frank Bretz²*

¹Gilead Sciences

²Novartis

dong.xil@gilead.com

The graphical approach to multiple testing provides a convenient tool for designing, visualizing, and performing multiplicity adjustments in confirmatory clinical trials while controlling the family-wise error rate. It assigns a set of weights to each intersection null hypothesis within the closed test framework. These weights form the basis for intersection tests using weighted individual p-values, such as the weighted Bonferroni test. However, the Hochberg procedure is not included in this framework, which limits its use in complex multiple test problems. In this talk, we introduce symmetric graphs to allow equal weights for the elementary null hypotheses associated with any intersection hypothesis so that the Hochberg procedure as well as omnibus tests such as Fisher's combination, O'Brien's, and F tests can be applied. We illustrate the method with a clinical trial example to show its benefits in visualizing complex multiple test problems.

A unified framework for weighted parametric group sequential design (WPGSD)

Keaven Anderson, Zifang Guo, Jing Zhao and ♦*Linda Sun*

Merck & Co., Inc.
linda.sun@merck.com

Contemporary clinical trials are getting more complex and usually with multiple primary objectives. Multiple primary objectives resulting in tests with known correlations include evaluating 1) multiple experimental treatment arms, 2) multiple populations, 3) the

combination of multiple arms and multiple populations. Group sequential design (GSD) is widely used in such clinical trials in which correlated tests of multiple hypotheses are used. In this presentation, we extend the framework of the weighted parametric multiple test procedure from fixed designs with a single analysis per objective to a GSD setting where different objectives may be assessed at the same or different times, each in a group sequential fashion. Pragmatic methods for design and analysis of weighted parametric group sequential design (WPGSD) under closed testing procedures are proposed to maintain the strong control of family-wise Type I error rate (FWER) when correlations between tests are incorporated. This results in the ability to relax testing bounds compared to designs not fully adjusting for known correlations, increasing power or allowing decreased sample size. We illustrate the proposed methods using clinical trial examples and conduct a simulation study to evaluate the operating characteristics.

Statistical Interactions in a Clinical Trial

Naitee Ting

Boehringer Ingelheim Pharmaceuticals, Inc.
naitee.ting@boehringer-ingelheim.com

Drug approval is a confirmatory practice - FDA needs to make a decision of approving or not approving a new drug. This decision is based on statistical hypothesis testing. When this is the case, alpha protection is critical. Primary statistical analysis needs to be clearly pre-specified - primary endpoint, primary comparison, and primary model. There should be no possibility for model selection. When statistical interaction presents, it is difficult to interpret treatment effect. Hence statistical interaction is not included in the primary model - it is estimated after clinical data read out.

Deep historical borrowing framework in confirmatory clinical trials with multiple endpoints

Tianyu Zhan

AbbVie
tianyu.zhan@abbvie.com

In current clinical trial development, historical information is receiving more attention as it provides utility beyond sample size calculation. Meta-analytic-predictive (MAP) priors and robust MAP priors have been proposed for prospectively borrowing historical data on a single endpoint. To simultaneously synthesize control information from multiple endpoints in confirmatory clinical trials, we propose to approximate posterior probabilities from a Bayesian hierarchical model and estimate critical values by deep learning to construct pre-specified strategies for hypothesis testing. This feature is important to ensure study integrity by establishing prospective decision functions before the trial conduct. Simulations are performed to show that our method properly controls family-wise error rate (FWER) and preserves power as compared with a typical practice of choosing constant critical values given a subset of null space. Satisfactory performance under prior-data conflict is also demonstrated. We further illustrate our method using a case study in Immunology.

Session 8H : Some Popular Applications In Data Integration

Multivariate Global-Local Priors for Small Area Estimation

Tamal Ghosh¹, ♦Malay Ghosh² and Jerry Maples³

¹Citibank, Tampa

²University of Florida

³United States Census Bureau

ghoshm@ufl.edu

It is now widely recognized that small area estimation (SAE) needs to be model-based. Global-local (GL) shrinkage priors for the random effects are important in sparse situations where many areas do not have any significant random effects. We propose in this paper a hierarchical multivariate model with GL priors. Such priors are useful when a multidimensional random effect vector has sparsity. We prove propriety of the posterior density when the regression coefficient matrix has improper uniform prior. Some concentration inequalities are derived for the tail probabilities of the shrinkage estimators.

Pseudo-Bayesian Small Area Estimation

♦ *Juhyung Lee¹, Gauri Datta² and Jiacheng Li²*

¹University of Florida

²University of Georgia
juhyunglee@ufl.edu

In a sample survey, a subpopulation is referred to as a "small area" or "small domain" if its sample is not large enough to yield alone an estimate of adequate precision of the subpopulation mean. The main interest in small area estimation is estimation of means of small areas. The observed best prediction (OBP) is a model-based prediction procedure for small area means that is proposed as an alternative to the empirical best linear unbiased prediction (EBLUP). The OBP method proposes an objective function to estimate the model parameters. We use this objective function to construct a pseudo likelihood for the model parameters. Using this likelihood we propose pseudo-Bayesian estimates of small area means under the Fay-Herriot model. Real data examples and simulation studies show that the pseudo-Bayesian estimators (PBE) compete favorably with the OBP and the EBLUPs. In terms of interval estimation, the PBE credible intervals attain the nominal coverage probability, while the OBP confidence intervals exhibit unsatisfactory coverage. Our simulations to investigate the robustness of the various predictors to mean misspecification show that the PBE predictors retain all the robustness enjoyed by the EBLUP and the OBP predictors. Simulations and data analysis show that the PBE predictors enjoy competitive frequentist properties. Being Bayesian by construction, the proposed PBE predictors admit a dual justification as optimal predictors, and they are expected to be attractive to practitioners.

Incorporating heterogeneous offsets in hierarchical disease mapping

♦ *Emily Peterson and Lance Waller*

Emory University
emily.nancy.peterson@emory.edu

Hierarchical spatial small area estimation of local disease rates has a rich history in spatial epidemiology and in spatial statistics. Generally, the hierarchical structure explores uncertainty in the small local counts of rare disease, the numerator of the local rate of disease, while the local population size, the denominator, is considered fixed. Data analysts now have multiple data products relating to local population size available, each with different errors. Here, we develop a hierarchical approach to combine three population data products from the U.S. Census, the decennial census, annual 5-year estimates from the rolling American Community Survey, and demographic estimates based on vital statistics. Each data product provides county-level estimates but with different error structures. We illustrate our approach on the 159 counties in the state of Georgia, United States, illustrating how the method adapts to the differing variance structures.

Session 9A : Bayesian Calibration Of Computer Models

On estimating photometric redshift of galaxies by augmenting observation with simulation

Arindam Fadikar

Argonne National Laboratory
afadikar@anl.gov

We are interested in obtaining a full probabilistic description of the redshift for cosmological objects based on photometric surveys. However, such surveys are limited and do not span the entire color spectrum which creates gaps in the input space for any statistical model to effectively reproduce the entire response surface. To overcome this challenge, the problem of redshift estimation is posed as an inverse problem where the observed photometric surveys are augmented with a simulation that, when calibrated, fills in the gaps in the prior training data and aids in quantifying the uncertainty in redshift estimation. We will discuss the emulation and calibration of this functional input–function output simulator and how such simulation augmentation technique improves uncertainty quantification in photometric redshift estimation.

A theoretical framework of the scaled Gaussian stochastic process in prediction and calibration

♦ *Mengyang Gu¹, Fangzheng Xie² and Long Wang³*

¹University of California, Santa Barbara

²Indiana University Bloomington

³Johns Hopkins University
mengyang@pstat.ucsb.edu

Model calibration or data inversion is one of the fundamental tasks in uncertainty quantification. In this work, we study the theoretical properties of the scaled Gaussian stochastic process (S-GaSP) for modeling the discrepancy between the reality and the imperfect mathematical model. We establish explicit connection between Gaussian stochastic process (GaSP) and S-GaSP through the orthogonal series representation. The predictive mean estimator in the S-GaSP calibration model converges to the reality at the same rate as the GaSP with a suitable choice of the regularization and scaling parameters. We also show that the calibrated mathematical model in the S-GaSP calibration converges to the one that minimizes the L_2 loss between the reality and the mathematical model, whereas the GaSP model with other widely used covariance functions does not have this property. Numerical examples confirm the excellent finite sample performance of our approaches.

Bayesian Calibration and Model Mixing

Matthew Pratola

Dept. of Statistics, The Ohio State University
mpratola@stat.osu.edu

Bayesian calibration has been a successful tool for combining evaluations of expensive simulators of theoretical models with observational data to construct improved predictions and estimate model parameters. However, the case of multiple simulators, each representing possibly unique theoretical models of the same phenomena, has received less attention. Meanwhile, Bayesian model averaging (BMA) is a popular tool for combining multiple statistical models by learning scalar model weights, which combine the available models in some optimal way. However, the scalar weights inherently assume a single overall best model while excluding the possibility of locally optimal models. An interesting case arises in certain physics models of nuclear dynamics, where locally optimal models are expected, and the spectrum of models can themselves be thought of as being indexed by calibration parameters. In this setting, one can draw a connection between Bayesian calibration and

BMA, which in turn motivates new statistical methodology for such physics problems involving multiple models.

An efficient approach for computer model calibration with variational Bayesian inference

♦ *Vojtech Kejzlar*¹ and *Taps Maiti*²

¹Skidmore College

²Michigan State University
vkejzlar@skidmore.edu

With the advancements in computer architectures, the use of computational models proliferates to solve complex problems in many scientific applications such as nuclear physics and climate research. However, the potential of such models is often hindered because they tend to be computationally expensive and consequently ill-fitting for uncertainty quantification. Furthermore, they are usually not calibrated with real-time observations. We present a computationally efficient algorithm based on variational Bayes inference (VBI) for the calibration of computer models with Gaussian processes. We provide empirical evidence for the computational scalability of our methodology and describe all the necessary details for an efficient implementation of the proposed algorithm. We also demonstrate the opportunities given by our method for practitioners on a real data nuclear physics application.

Session 9B : Novel Developments For Functional Data Analysis

Adaptive Frequency Band Analysis for Functional Time Series

♦ *Pramita Bagchi*¹ and *Scott Bruce*²

¹George Mason University

²Texas A&M University
pbagchi@gmu.edu

The frequency-domain properties of nonstationary functional time series often contain valuable information. These properties are characterized through their time-varying power spectrum. Practitioners seeking low-dimensional summary measures of the power spectrum often partition frequencies into bands and create collapsed power measures within bands. However, standard frequency bands have primarily been developed through manual inspection of time series data and may not adequately summarize power spectra. In this article, we propose a framework for adaptive frequency band estimation of nonstationary functional time series that optimally summarizes the time-varying dynamics of the series. We develop a scan statistic and search algorithm to detect changes in the frequency domain. We establish the theoretical properties of this framework and develop a computationally-efficient implementation. The validity of our method is also justified through numerous simulation studies and an application to analyzing electroencephalogram data in participants alternating between eyes open and eyes closed conditions.

Sliced Elastic Distance for Climate Model Validation

*Robert Garrett*¹, ♦ *Trevor Harris*² and *Bo Li*¹

¹University of Illinois at Urbana-Champaign

²Texas A&M University
tharris@stat.tamu.edu

Global climate model validation is integral for ensuring climate models produce realistic climatologies. However, many post-hoc statistical evaluation methods rely on simplifying models that discard information and fail to distinguish between different sources of variability. Here, we introduce a functional data analysis approach for computing sliced amplitude and phase distances between spatiotemporal processes, analogous to the sliced Wasserstein distance.

Because our method uses time-warping, which respects temporal ordering, we can more precisely quantify differences between climate models than the previous Wasserstein-based approach. Finally, we apply our method to compare the performance of CMIP5 vs. CMIP6 models in representing historical surface temperature and precipitation from 1979-2005.

MARGARITA: Marginal-Product Basis Representation for Multi-dimensional Functional Data Analysis

William Consagra, Arun Venkataraman and ♦ Xing Qiu

University of Rochester
xing_qiu@urmc.rochester.edu

Modern neuroimaging data can be modeled as noisy observations of an underlying multidimensional random spatial or spatio-temporal function. Many traditional techniques from functional data analysis are plagued by the curse of dimensionality and quickly become intractable as the dimension of the domain increases. In this talk, we propose a framework for learning multidimensional continuous representations of neuroimages that is immune to several manifestations of the curse. Our representations are defined to be multiplicatively separable and adapted to the data according to a criterion similar to a multidimensional functional principal components analysis. We show that the resulting estimation problem can be solved efficiently by the tensor decomposition of a carefully defined reduction transformation of the observed data. The incorporation of both regularization and dimensionality reduction is discussed. The advantages of the proposed method over competing methods are demonstrated with both simulated data and a set of diffusion tensor imaging data collected from subjects with traumatic brain injury.

Session 9C: Statistical Methods For High Dimensional Microbiome Data

What Can We Learn About the Bias of Microbiome Studies from Analyzing Data from Mock Communities

*Mo Li*¹, ♦ *Glen Satten*², *Ni Zhao*¹, *Angel Rivera*³ and *Robert Tyx*³

¹Johns Hopkins University

²Emory University

³CDC
gsatten@emory.edu

It is known that data from both 16S and shotgun metagenomics studies are subject to biases that cause the observed relative abundances of taxa to differ from their true values. Mock community analyses, in which the relative abundances of all taxa in the sample are known by construction, seem to offer the hope that these biases can be measured. However, it is unclear whether the bias we measure in a mock community analysis is the same as we measure in a sample in which taxa are spiked in at known relative abundance, or if the biases we measure in spike-in samples is the same as the bias we would measure in a real (e.g., biological) sample. Here we consider these questions in the context of 16S rRNA measurements on three sets of samples: the commercially-available Zymo cells mock community; the Zymo mock community mixed with Swedish Snus, a smokeless tobacco product that is virtually bacteria-free; and a set of commercially-available smokeless tobacco products. Each set of samples was subject to four different extraction protocols. The goal of our analysis is to determine whether the patterns of bias observed in each set of samples is the same, i.e. can we learn about the bias in the commercially-available smokeless tobacco products by studying the Zymo cells mock community.

Nonparametric Bayesian approaches for identifying differentially abundant genera between multiple groups in microbiome data

Archie Sachdeva, Somnath Datta and [♦]Subharup Guha

University of Florida
s.guha@ufl.edu

We propose nonparametric Bayesian methods for identifying microbiomic signatures consisting of differentially abundant genera between various subgroups of a population and studying their association with demographic and clinicopathological attributes. The new statistical approaches combine the sparsity-inducing property of Dirichlet processes to detect optimally sparse non-linear combinations of high dimensional microbiome abundances and subject characteristics. Simulation studies demonstrate the benefits of the approach relative to existing methods. The technique is applied to detect differential taxa among adults and children in an oral microbiomic data set with detailed subject-specific information such as biological sex, BMI, beverage intake, oral health practices, and family relationships.

Deep ensemble learning over the microbial phylogenetic tree (DeepEn-Phy)

[♦]Wodan Ling¹, Youran Qi², Xing Hua¹ and Michael Wu¹

¹Fred Hutchinson Cancer Center

²Amazon
wling@fredhutch.org

Successful prediction of clinical outcomes facilitates tailored diagnosis and treatment. The microbiome has been shown to be an important biomarker to predict host clinical outcomes. Further, the incorporation of microbial phylogeny, the evolutionary relationship among microbes, has been demonstrated to improve prediction accuracy. We propose a phylogeny-driven deep neural network (PhyNN) and develop an ensemble method, DeepEn-Phy, for host clinical outcome prediction. The method is designed to optimally extract features from phylogeny, thereby take full advantage of the information in phylogeny while harnessing the core principles of phylogeny (in contrast to taxonomy). We apply DeepEn-Phy to a real large microbiome data set to predict both categorical and continuous clinical outcomes. DeepEn-Phy demonstrates superior prediction performance to existing machine learning and deep learning approaches. Overall, DeepEn-Phy provides a new strategy for designing deep neural network architectures within the context of phylogeny-constrained microbiome data.

IFAA: Robust association identification and Inference For Absolute Abundance in microbiome analyses

Zhigang Li

University of Florida
zhigang.li@ufl.edu

The target of inference in microbiome analyses is usually relative abundance (RA) because RA in a sample (e.g., stool) can be considered as an approximation of RA in an entire ecosystem (e.g., gut). However, inference on RA suffers from the fact that RA are calculated by dividing absolute abundances (AA) over the common denominator (CD), the summation of all AA (i.e., library size). Because of that, perturbation in one taxon will result in a change in the CD and thus cause false changes in RA of all other taxa, and those false changes could lead to false positive/negative findings. We propose a novel analysis approach (IFAA) to make robust inference on AA of an ecosystem that can circumvent the issues induced by the CD problem and compositional structure of RA. IFAA can also address the issues of overdispersion and handle zero-inflated data structures. IFAA identifies microbial taxa associated with the

covariates in Phase one and estimates the association parameters by employing an independent reference taxon in Phase two. Two real data applications are presented and extensive simulations show that IFAA outperforms other established existing approaches by a big margin in the presence of unbalanced library size.

Session 9D : Recent Advancements In Statistical Data Integration

Meta Clustering for Collaborative Learning

[♦]Chenglong Ye¹, Reza Ghanadan² and Jie Ding³

¹University of Kentucky

²Google

³University of Minnesota

chenglong.ye@uky.edu

An emerging number of learning scenarios involve a set of learners/analysts each equipped with a unique dataset and algorithm, who may collaborate with each other to enhance their learning performance. From the perspective of a particular learner, a careless collaboration with task-irrelevant other learners is likely to incur modeling error. A crucial problem is to search for the most appropriate collaborators so that their data and modeling resources can be effectively leveraged. Motivated by this, we propose to study the problem of 'meta clustering', where the goal is to identify subsets of relevant learners whose collaboration will improve the performance of each individual learner. In particular, we study the scenario where each learner is performing a supervised regression, and the meta clustering aims to categorize the underlying supervised relations (between responses and predictors) instead of the raw data. We propose a general method named as Select-Exchange-Cluster (SEC) for performing such a clustering. Our method is computationally efficient as it does not require each learner to exchange their raw data. We prove that the SEC method can accurately cluster the learners into appropriate collaboration sets according to their underlying regression functions. Synthetic and real data examples show the desired performance and wide applicability of SEC to a variety of learning tasks.

Joint integrative analysis of dependent data sources

[♦]Emily Hector¹ and Peter Song²

¹North Carolina State University

²University of Michigan

ehector@ncsu.edu

We propose a distributed quadratic inference function framework to jointly estimate regression parameters from multiple potentially heterogeneous data sources with correlated vector outcomes. The primary goal of this joint integrative analysis is to estimate covariate effects on all outcomes through a marginal regression model in a statistically and computationally efficient way. We develop a data integration procedure for statistical estimation and inference of regression parameters that is implemented in a fully distributed and parallelized computational scheme. To overcome computational and modeling challenges arising from the high-dimensional likelihood of the correlated vector outcomes, we propose to analyze each data source using quadratic inference functions, and then to jointly reestimate parameters from each data source by accounting for correlation between data sources using a combined meta-estimator in a similar spirit to generalized method of moments. We show both theoretically and numerically that the proposed method yields efficiency improvements and is computationally fast. We illustrate the

proposed methodology with the joint integrative analysis of the association between smoking and metabolites in a large multi-cohort study.

Data Integration Via Analysis of Subspaces

♦ *Jack Prothero*¹, *Meilei Jiang*², *Quoc Tran-Dinh*³, *Jan Hammig*³ and *J.s. Marron*³

¹National Institute of Standards and Technology

²Meta

³UNC Chapel Hill

jack.prothero@nist.gov

Modern data collection in bioinformatics and other big-data paradigms often incorporates traits derived from multiple different points of view of the observations. We call this data multi-view or multi-block data. The emergent field of data integration develops and applies new methods for studying multi-block data and identifying how different data blocks relate and differ. One major frontier in contemporary data integration research is methodology that can identify partially-shared structure between sub-collections of data blocks. This work presents our new method on this frontier: Data Integration Via Analysis of Subspaces (DIVAS). DIVAS combines new insights in angular subspace perturbation theory with recent developments in matrix signal processing and convex-concave optimization into one algorithm for parsing partially shared structure. Our novel approach based on principal angles between subspaces provides built-in inference on the results of the analysis, and is effective even in high-dimension-low-sample-size (HDLSS) situations.

Synthetic-data-based transfer learning approach for multi-site risk prediction

♦ *Tian Gu* and *Rui Duan*

Department of Biostatistics, Harvard T.H. Chan School of Public Health

tiangu@hsph.harvard.edu

We propose a synthetic-data-based transfer learning approach (SynTL) to incorporate multi-site healthcare data for improving the risk prediction of a target population, accounting for challenges including heterogeneity, data sharing, and privacy constraints. SynTL combines locally trained risk prediction models from each source population with the target population to adjust for the data heterogeneity through a flexible distance-based transfer learning approach. Heterogeneity-adjusted synthetic data are generated for source populations where individual-level data are not shareable. The synthetic data are then combined with the target and source data for joint training of the target model. We evaluate SynTL via extensive simulation studies and an application to multi-site data from the electronic Medical Records and Genomics (eMERGE) Network to predict the risk of extreme obesity. SynTL is shown to improve the risk prediction performance of the target population, and is robust to the level of heterogeneity between the target and source populations. It protects patient-level information and is highly communication efficient.

Session 9E : Modern Business Statistical Analysis

Penalized quantile regression

♦ *Ben Sherwood* and *Shaobo Li*

University of Kansas

ben.sherwood@ku.edu

Quantile regression directly models a conditional quantile. Penalized quantile regression constrains the regression coefficients simi-

lar to penalized mean regression. Quantile regression with a lasso penalty can be reframed as a quantile regression problem with augmented data and therefore can be formulated as a linear programming problem. If a group lasso penalty is used, then it becomes a second order cone programming problem. These approaches become computationally burdensome for large values of n or p . Using a Huber approximation to the quantile function allows for the use of computationally efficient algorithms that require a differentiable loss function that can be implemented for both penalties. These algorithms then can be used as the backbones for implanting penalized quantile regression with other penalties such as Adaptive Lasso, SCAD, MCP and group versions of these penalties.

On the use of Minimum Penalties in Multivariate Regression

♦ *Brad Price*¹ and *Ben Sherwood*²

¹West Virginia University

²University of Kansas

brad.price@mail.wvu.edu

Modern multivariate machine learning and statistical methodologies estimate parameters of interest while leveraging prior knowledge of the association between outcome variables. The methods that do allow for estimation of relationships do so typically through an error covariance matrix in multivariate regression which does not scale to other types of models. In this article we proposed the MinPEN framework to simultaneously estimate regression coefficients associated with the multivariate regression model and the relationships between outcome variables using mild assumptions. The MinPen framework utilizes a novel penalty based on the minimum function to exploit detected relationships between responses. An iterative algorithm that generalizes current state of the art methods is proposed as a solution to the non-convex optimization that is required to obtain estimates. Theoretical results such as high dimensional convergence rates, model selection consistency, and a framework for post selection inference are provided. We extend the proposed MinPen framework to other exponential family loss functions, with a specific focus on multiple binomial responses. Tuning parameter selection is also addressed. Finally, simulations and two data examples are presented to show the finite sample properties of this framework. This is joint work with Ben Sherwood from the University of Kansas.

Joint Modeling of Playing Time and Purchase Propensity in Massively Multiplayer Online Role Playing Games Using Crossed Random Effects

Trambak Banerjee

University of Kansas

trambak@ku.edu

Massively Multiplayer Online Role Playing Games (MMORPGs) offer a unique blend of a personalized gaming experience and a platform for forging social connections. Managers of these digital products rely on predictions of key player responses, such as playing time and purchase propensity, to design timely interventions for promoting, engaging and monetizing their playing base. However, the longitudinal data associated with these MMORPGs not only exhibit a large set of potential predictors to choose from but often present several other distinctive characteristics that pose significant challenges in developing flexible statistical algorithms that can generate efficient predictions of future player activities. For instance, the existence of virtual communities or 'guilds' in these games complicate prediction since players who are part of the same guild have correlated behaviors and the guilds themselves evolve over time and, thus, have a dynamic effect on the future playing behavior of its

members. In this paper, we develop a Crossed Random Effects Joint Modeling (CREJM) framework for analyzing correlated player responses in MMORPGs. Contrary to existing methods that assume player independence, CREJM is flexible enough to incorporate both player dependence as well as time varying guild effects on the future playing behavior of the guild members. On a large-scale data from a popular MMORPG, CREJM conducts simultaneous selection of fixed and random effects in high-dimensional penalized multivariate mixed models. We study the asymptotic properties of the variable selection procedure in CREJM and establish its selection consistency. Besides providing superior predictions of daily playing time and purchase propensity over competing methods, CREJM also predicts player correlations within each guild which are valuable for optimizing future promotional and reward policies for these virtual communities.

Measuring goodness-of-fit for bankruptcy prediction and its application to U.S. and Polish data

♦Xiaorui Zhu and Dungan Liu

University of Cincinnati
zhuxr@mail.uc.edu

The binary and ordinal outcomes are common in economic, political, behavioral, and psychological studies. However, the ordinal nature of rank data presents special challenges in statistical inference, and many statistical problems remain unsolved. To make inferences for the ordinal data, especially in finance, the logit model is one of the most widely used generalized linear models. But two crucial voids have been sought to addressed. First, there is no efficient tool for assessing the explanatory power of each determinant. Second, existing pseudo- R^2 cannot resemble the ordinary least square (OLS) R^2 that represents the adequacy of prediction models. To address these voids, we propose a new goodness-of-fit measure for the logit model and apply it to the U.S. and Polish bankruptcy data. Specifically, we simulate a continuous variable S as a surrogate of the bankruptcy likelihood (Liu and Zhang, 2018). Then the proposed R^2 represents the proportion of the variance of the surrogate response explained by explanatory variables through a linear model, and we call it a surrogate R^2 . Simulation and the real data results demonstrate our surrogate R^2 can approximate the ordinary least square (OLS) R^2 used for linear models. Besides, it can quantify the explanatory power of an individual or a set of predictors.

Session 9F : Application And Theory Of Statistical Test And Evaluation

Statistical Evaluation of Deep Learning-based Side-channel Analysis

Aidong Ding

Northeastern University
a.ding@northeastern.edu

Side-channel attacks (SCAs) use physical side-channel measurements, such as power consumption or the execution time, to break theoretical secure cryptographic algorithms in the implemented systems. Deep Learning (DL)-based SCAs recently were able to break cryptographic implementations protected with countermeasures against traditional template attacks. We study statistical models to evaluate risks of SCA. An efficient algorithm for calculating Gaussian Entropy (GE) is proposed that can evaluate the effectiveness of DL-based SCAs as well as traditional SCAs. Training Deep Neural Networks (DNNs) with traditional loss functions often do not lead to optimal/successful follow-on DL-based SCAs. Apply-

ing the new GE estimation, the DNNs can be trained to optimize for follow-on SCA performance.

Improved Meta-Analysis of ROC curves

♦Buddika Peiris¹ and Shuang Yang²

¹Worcester Polytechnic Institute

²Worcester Polytechnic Institute
tbpeiris@wpi.edu

Improved Meta-Analysis in ROC Curves Buddika Peiris Worcester Polytechnic Institute, Worcester MAAbstract In this work we propose an improved meta-analysis method to synthesize the ROC curves from multiple individual studies. When synthesizing, ROC curves are transformed to linear models then combined and re-transformed to a ROC curve. The existing method proposed in Kester and Buntinx (2000) has an issue since information about the covariance between the coefficients of the same model and covariance among coefficients across studies are not utilized. Here we introduce how to take the correlations between the coefficients and the correlations between studies into account to improve the existing meta-analysis. We provide numerical example to compare the proposed methods with the existing method by using mean square prediction errors with applications to forecasting problem in Environmental study. Keywords: ROC Curve, meta-analysis, generalized least square, Der Simonian and Laird, synthesis of slopes.

Signal-noise ratio of genetic associations and statistical power of SNP-set tests

Hong Zhang¹, Ming Liu², Jiashun Jin³ and ♦Zheyang Wu²

¹Merck Research Laboratories

²WPI

³Carnegie Mellon University
zheyangwu@wpi.edu

The SNP-set analysis is a powerful tool for dissecting the genetics of complex human diseases. There are three fundamental genetic association approaches to SNR-set analysis: the marginal model fitting approach, the joint model fitting approach, and the decorrelation approach. A problem of primary interest is how these approaches compare with each other. To address this problem, we develop a theoretical platform to compare the signal-to-noise ratio (SNR) of these approaches under the generalized linear model. We elaborate how causal genetic effects give rise to statistically detectable association signals, and show that when causal effects spread over blocks of strong linkage disequilibrium (LD), the SNR of the marginal model fitting is usually higher than that of the decorrelation approach, which in turn is higher than that of the unbiased joint model fitting approach. We also scrutinize dense effects and LDs by a bivariate model and extensive simulations using the 1000 Genome Project data. Last, we compare the statistical power of two generic types of SNP-set tests (summation-based and supremum-based) by simulations and an osteoporosis study using large data from UK Biobank. Our results help develop powerful tools for SNP-set analysis and understand the signal detection problem in the presence of colored noise.

BEAUTY Powered BEAST

Kai Zhang¹, ♦Zhigen Zhao² and Wen Zhou³

¹UNC

²Temple

³Colorado State University
zhaozhg@temple.edu

We study nonparametric dependence detection with the proposed binary expansion approximation of uniformity (BEAUTY) approach, which generalizes the celebrated Euler's formula, and ap-

proximates the characteristic function of any copula with a linear combination of expectations of binary interactions from marginal binary expansions. This novel theory enables a unification of many important tests through approximations from some quadratic forms of symmetry statistics, where the deterministic weight matrix characterizes the power properties of each test. To achieve a robust power, we study test statistics with data-adaptive weights, referred to as the binary expansion adaptive symmetry test (BEAST). By utilizing the properties of the binary expansion filtration, we show that the Neyman-Pearson test of uniformity can be approximated by an oracle weighted sum of symmetry statistics. The BEAST with this oracle provides a benchmark of feasible power against any alternative by leading all existing tests with a substantial margin. To approach this oracle power, we develop the BEAST through a regularized resampling approximation of the oracle test. The BEAST improves the empirical power of many existing tests against a wide spectrum of common alternatives and provides clear interpretation of the form of dependency when significant.

Session 9G : Statistical Challenges In Clinical Trials For Alzheimer Disease

Dose change and statistical power in the Aducanumab trial

Guogen Shan

University of Florida
gshan@ufl.edu

The search for disease modifying therapies in Alzheimer's disease (AD) has recently led to promising results but also revealed design issues in the clinical trials themselves. One of the statistical challenges raised from contemporary AD trials is the dose change after an interim analysis. This research was motivated by the very recent Aducanumab trial to study the effects of dose change on statistical power of an AD trial when dose change occurs. We conduct extensive simulation studies to calculate statistical powers when the relationship between treatment effect size and time is linear non-linear, and the investigated drug has delayed treatment effect or not. Statistical powers depend on many design factors including the dose change time, correlation, population homogeneity, and treatment effect time. We would recommend researchers conduct sufficient enough simulation studies at the interim analysis to justify the modification

ed sample size and/or follow-up time modification meanwhile the type I and II error rates are controlled.

A More Efficient Outcome for Alzheimer Disease Research: the Item Response Theory Based Score for the Clinical Dementia Rating (CDRR)

♦ *Yan Li, Guoqiao Wang, Chengjie Xiong, Krista L Moulder and John C Morris*

Washington University in St. Louis
yanli833@wustl.edu

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 sever-

ity. 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 (ζ 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 ζ 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).

Session 9H : Statistics Education In The Era Of Ai And Data Science

Interactive Graphics: A Bridge from Coding to Programming

Adam Loy

Carleton College
aloy@carleton.edu

Ten years ago, Nolan and Temple Lang argued for the growing importance of computational and technological emphasis in the statistics curricula. Their call catalyzed the rapid changes seen in curricular focus on data handling and computation with real data in statistics, as well as the rise of data science education. However, we feel Nolan and Temple Lang's call did not lead to sufficient change in the teaching of data visualization and web-based technologies. While statistics educators have vastly increased their use of static graphics in the curricula, there is a pressing need for more exposure to interactive graphics. The ability to create dynamic and interactive visualizations is an in-demand professional skill for technical communication. In addition, these visualizations inherently reinforce multivariate thinking, broaden the concepts of statistical computing, deepen computational reasoning, and allow students to compute with data in the practice of statistics. In this talk, I will outline why dynamic and interactive graphics should be incorporated in the statistics curricula and how to use interactive graphics via web-based apps to discuss key programming ideas, such as the modularization of code.

Case studies to community engagement: bringing hands-on data science experiences to the classroom

♦ *Carrie Wright¹, Stephanie Hicks¹, Ava Hoffman¹, Michael Rosenblum¹, Michael Breshock¹, Qier Meng¹, Margaret Taub, Leah Jager¹, Tyler Derreth¹ and Mindi Levin¹*

¹Johns Hopkins Bloomberg School of Public Health

cwrigh60@jhu.edu

Research suggests that experiential learning is more conducive for lasting memory, is more engaging, and helps learners to apply their knowledge in different contexts. In particular when it comes to data science, although simple toy datasets can be best for introducing topics, giving learners an awareness of the context and complexities of more realistic analyses better prepares them for the work they may encounter in research or on the job. Furthermore, providing students with data science and statistical lessons within a framework focused on reproducibility, transparency, and data ethics, provides students with a more cohesive understanding of how to apply data science methods in a responsible and mindful manner. To provide learners with such opportunities we have created two initiatives. The first is called the Open Case Studies project (<https://www.opencasestudies.org/>), which aims to provide learners and educators with a library of complete and peer-reviewed data analysis guides using publicly available data on various statistical, data science, and public health topics. The intention is for these case studies to be used within and outside of the classroom to help learners understand the decision making process and the various data science and statistical steps involved in data analysis. Our educator guide (https://www.opencasestudies.org/OCS_Guide/) can help instructors implement our case studies in the classroom. The second initiative is a course (https://jhudatascience.org/Baltimore_Community_Course/) for critical service-learning data science projects with local nonprofit organizations. This course gives students a sense of the realities of working with real-world data, exposure to the challenges of communicating with others with different perspectives, experience working through data ethics challenges, and an appreciation for how domain expertise can better inform data science work. We plan to continue to assess the impact of these initiatives and hope to expand these projects to reach more learners and educators.

Constructing a Modern Data Visualization Course: Topics, Reflections, and Feedback

Steven Foti

University of Florida
fotisj@ufl.edu

Creating high quality visualizations is an important skill for students from all backgrounds and academic programs that work with data

in any capacity. Whether they are analyzing data, interpreting data, or sharing the results of their research to diverse audiences across multiple media formats, visualizations provide an efficient way to convey information. This presentation will provide an overview of a data visualization course that is intended for graduate students in both biostatistics and the broader health sciences. The discussion will emphasize course topics that weave fundamental ideas in graphing with coding skills in R, project ideas that connect students from multiple disciplines to relevant issues in the community, and a reflection of experiences from three years of teaching the course. Participants are encouraged to engage in the conversation and share any feedback they may have.

Foundations for NLP-assisted formative assessment feedback for short-answer tasks in large-enrollment classes

Susan Lloyd, ♦Matthew Beckman, Dennis Pearl, Rebecca Pas-soneau, Zhaohui Li and Zekun Wang

Penn State University
mdb268@psu.edu

Research has linked "write-to-learn" tasks to improved learning outcomes, yet constructed-response methods of formative assessment become unwieldy for instructors with large class sizes. Responses collected on six short-answer tasks, from a sample of 1,935 students, were divided among trained evaluators to measure inter- and intra-rater agreement. A natural language processing (NLP) algorithm scored a subset of student responses for correctness. Quadratic weighted kappas (QWK) for pairwise rater agreement between 0.74 - 0.83, and Fleiss' Kappa of 0.68 for group agreement, indicate substantial inter-rater agreement among raters, including the algorithm. One rater had reviewed a subset of 188 responses seven years earlier, resulting in an estimated intra-rater agreement of QWK = 0.89. Two NLP algorithms were compared, the first being a logistic regression (LR) combined with a Long Short-Term Memory (LSTM) for learning vector representations, and the second being the Semantic Feature-Wise Transformation Relation Network (SFRN). The team then piloted cluster analysis of student responses to determine if a set of student responses that have the same correctness can be grouped into semantically similar clusters. Future work will refine the clustering algorithm such that instructors may interpret meanings common within clusters to provide scalable, formative assessment feedback.

Index of Authors

- Albert, P, 57, 103, 104
Alemdjrodo, K, **57, 101**
Alsharman, M, **52, 83**
Anderson, K, 60, 113
Aron, J, **57, 103**
Avilamatos, L, 59, 110
- Bacher, R, **49, 71**
Bacher, RL, 49, 71
Baek, J, 49, 73
Bagchi, P, **61, 115**
Bai, R, **54, 92**
Baker, J, 50, 52, 78, 84
Bakitas, M, 52, 86
Bal, AB, 51, 82
Bandyopadhyay, D, 60, 112
Banerjee, S, 59, 109
Banerjee, T, **61, 117**
Banker, M, 49, **55, 73, 95**
Bazan, J, **56, 56, 98, 98**
Beckman, M, 62, 120
Belitskaya-Levy, I, 48, 69
Berger, J, 53, 86
Berry, S, 53, 86
Bhaduri, M, **53, 88**
Bhattacharya, I, **51, 82**
Bhattacharya, S, 47, 66
Bing, X, **47, 65**
Blei, D, 47, 65
Bonvini, M, **54, 90**
Bornkamp, B, 50, 75
Bradley, P, 56, 98
Breshock, M, 62, 119
Bretz, F, 60, 113
Bruce, S, 61, 115
Brunel, N, 56, 100
Buhlmann, P, 52, 85
Bunea, F, 47, 65
- Cai, C, 53, 88
Cai, G, **58, 105**
Cai, H, **52, 84**
Cai, W, 53, 89
Cai, X, 50, **52, 78, 84**
Candès, E, 52, 84
Cao, J, 51, 80
Cao, S, **58, 105**
Cao, X, **54, 92**
Cao, Y, 49, 72
- Carone, M, 58, 107
Carroll, R, 55, 93
Castellucci, L, 53, 86
Castro, LM, **59, 110**
Cevid, D, 52, 85
Chandrasena, S, **51, 79**
Chang, C, **50, 51, 57, 76, 80,**
101
Chang, V, **53, 89**
Chatterjee, N, **55**
Chen, A, 58, 104
Chen, C, 49, **50, 73, 77**
Chen, D, **52, 85**
Chen, H, 49, 73
Chen, I, 60, 111
Chen, J, **48, 49, 54, 60, 67,**
72, **93, 111**
Chen, K, **52, 83**
Chen, Q, 55, 94
Chen, S, 48, 50, 51, **56, 68,**
77, 81, **100**
Chen, T, **48, 48, 69, 69**
Chen, W, **57, 102**
Chen, X, 49, **54, 72, 90**
Chen, Y, **53, 53, 56, 87, 89,**
99
Chen, Z, 49, 72
Cheng, S, **58, 107**
Cheng, Y, 51, 79
Cheung, Li, 57, 103
Chiou, SH, 53, 86
Choo-Wosoba, H, **57, 104**
Chung, D, 54, 91
Chung', M, **51, 82**
Coffman, D, **47, 65**
Colditz, G, 51, 80
Consagra, W, 61, 115
Cook, D, 54, 56, 90, 97
Cook, RD, 54, 91
Coull, B, 50, 78
Cui, X, 56, 100
Cui, Y, 56, 97
Cukras, S, 58, 104
- Dai, B, **55, 96**
Dalmacy, D, 52, 83
Dang, X, 57, 102
Daniels, M, 52, 86
Das, S, 50, 77
- Dasilva, M, **57, 103**
Datta, G, 60, 114
Datta, S, 49, 61, 71, 116
Daver, N, 53, 88
Degruttola, V, 48, 69
Delrocco, N, **48, 70**
Delvin, B, 51, 81
Deng, K, 58, 106
Deng, S, **54, 54, 89, 93**
Derreth, T, 62, 119
Dewhirst, F, 50, 78
Diao, G, 57, 101
Digravio, C, 48, 70
Ding, A, 48, **61, 68, 70, 118**
Ding, J, 61, 116
Ding, S, **54, 91**
Dong, G, 53, 89
Doshi-Velez, F, 54, 91
Duan, L, **54, 92**
Duan, R, **53, 53, 61, 87, 87,**
117
Dunson, D, 50, 74
- Ertefaie, A, 51, 82
- Fadikar, A, **60, 114**
Fan, J, **53**
Fang, Y, 51, 80
Farkouh, M, 53, 86
Ferrall-Fairbanks, M, **56, 98**
Ferrarelli, F, 59, 109
Fletcher, PT, **51, 82**
Follmann, D, **58, 106**
Forzani, L, **54, 91**
Foti, S, **62, 120**
Fu, H, **55, 94**
Fuquenepatino, JA, **54, 92**
- G'sell, M, 51, 81
Galarza, C, 59, 110
Gao, Y, 59, 110
Garay, A, 56, 98
Garrett, R, **61, 115**
Ghanadan, R, 61, 116
Ghosh, M, 60, 113
Ghosh, T, **60, 113**
Giessing, A, **52, 84**
Gilbert, P, 55, 58, 93, 107
Goligher, E, 53, 86
- Gribble, M, 57, 103
Gu, M, **60, 114**
Gu, T, **61, 117**
Gu, Y, **50, 58, 74, 107**
Guha, S, **60, 61, 112, 116**
Guhaniyogi, R, **60, 60, 112,**
112
Gui, J, **60, 111**
Guinness, J, 50, 78
Guo, J, **48, 68**
Guo, S, 48, 67
Guo, X, 51, 82
Guo, Y, **52, 85**
Guo, Z, **52, 60, 85, 113**
- Habiger, J, 59, 108
Han, P, **53, 87**
Hannig, J, 61, 117
Hao, N, 55, 58, 93, 106
Harhay, MO, 54, 90
Harris, T, 61, 115
Harrison, C, **59, 108**
Hartigan, P, 52, 83
Hassan, T, 49, 74
He, Q, **59, 59, 108, 108**
He, W, 55, 93
He, X, 52, 85
He, Y, **52, 85**
Hector, E, **61, 116**
Hicks, S, 49, 62, 72, 119
Higgs, R, 56, 99
Ho, Y, 59, 108
Hoen, A, 60, 111
Hoffman, A, 62, 119
Hou, W, 49, 72
Hu, J, **56, 99**
Hua, X, 61, 116
Huang, C, 53, 86
Huang, H, **59, 59, 108, 108,**
109, 109
Huang, J, **53, 87**
Huang, Y, **58, 107**
Huey, N, 50, 78
Huo, X, **48, 68**
Huo, Z, **49, 49, 71, 71**
- Imperato, J, 49, 74
Jaeger, B, **54, 91**

- Jager, L, 62, 119
Jankar, J, **53, 89**
Janson, L, 58, 108
Jantre, S, **47, 66**
Jawahri, A, 52, 86
Jensen, E, 49, 73
Ji, H, **49, 72**
Ji, P, **57, 103**
Ji, Z, 49, 72
Jia, B, 56, 99
Jiang, M, **60, 61, 112, 117**
Jiang, S, **51, 80**
Jiang, Y, **58, 104**
Jin, J, **61, 118**
Jin, W, 49, 72
Jin, Y, **52, 55, 84, 95**
Johnson, B, 51, 82
Jones, G, 54, 92
Joshi, S, **52, 83**
- Kalantar-Zadeh, K, 59, 109
Kang, J, **48, 68**
Kang, Y, **49, 73**
Karmakar, B, 50, 77
Ke, H, 51, 81
Ke, Y, **57, 57, 103, 103**
Keele, LJ, 54, 90
Kejzlar, V, **60, 115**
Keko, M, 52, 83
Kennedy, EH, 54, 90
Kil, S, 50, 75
Kong, D, 51, 80
Kong, M, 50, 77
Kornblith, L, 53, 86
Krafty, R, **59, 109**
Kurum, E, **47, 59, 65, 109**
Kwizera, M, 55, 94
Kyriakides, TC, 52, 83
- Lachos, V, 56, 98
Lachos, VH, **59, 59, 110, 110**
Laha, N, **50, 78**
Lan, Z, 49, **51, 71, 79**
Langschumacher, F, 59, 110
Lawler, P, 53, 86
Lee, D, 53, 88
Lee, H, **51, 81**
Lee, J, 48, 57, **60, 69, 101, 114**
Lee, K, 54, 92
Lee, KH, 50, 78
Lee, S, 60, 112
Lee, Y, 50, 77
Leifer, E, **53, 86**
Levin, M, 62, 119
Li, B, **54, 56, 61, 90, 97, 115**
Li, C, 55, 96
Li, D, **51, 80**
Li, F, 54, 90
Li, G, 48, 69
Li, H, 57, 102
Li, J, **49, 50, 58, 60, 74, 75, 104, 114**
Li, JJ, **57, 102**
Li, M, 53, **54, 60, 61, 89, 93, 111, 115**
Li, N, 48, 69
Li, R, 52, 85
Li, S, 48, 61, 69, 117
Li, X, 56, **57, 99, 100**
Li, Y, **52, 59, 62, 85, 109, 119**
Li, Z, 52, 59, 60, **61, 62, 86, 109, 111, 112, 116, 120**
Liang, F, **47, 66**
Liang, M, **50, 76**
Lin, Q, 51, 79
Lin, R, **57, 100**
Lin, X, **55, 55, 95**
Ling, W, **61, 116**
Little, R, 55, 94
Liu, D, 61, 118
Liu, J, **55, 94**
Liu, L, 52, 85
Liu, M, 55, 61, 94, 118
Liu, P, **51, 81**
Liu, R, 51, 79
Liu, S, **56, 98**
Liu, W, **52, 85**
Liu, Y, 50, **56, 75, 99**
Lloyd, S, **62, 120**
Long, Q, 50, 52, 76, 77, 83
Loorvaleriano, KA, **59, 110**
Lotspeich, S, **48, 71**
Lou, Y, **55, 94**
Loy, A, **62, 119**
Loyal, J, **56, 97**
Lu, J, 54, 55, 89, 95
Lu, M, **49, 72**
Lu, Q, 54, 93
Lu, W, 51, 52, 80, 84
Lu, Y, 48, 69
Lu, Z, 56, 99
Luo, X, **50, 58, 75, 105**
Lyu, T, **50, 75**
- Ma, C, 59, 111
Ma, J, 53, 88
Ma, P, **56, 100**
Ma, S, 52, 83
Ma, T, 51, 81
Ma, X, 53, **55, 86, 96**
Mackey, S, 47, 65
Madireddy, S, **47, 67**
Mai, Q, **58, 106**
Maiti, T, 47, 60, 66, 115
Majumder, S, **50, 78**
Mallick, B, 60, 113
Mandal, A, **53, 53, 89, 89**
Maples, J, 60, 113
Markwelch, J, 50, 78
Maronge, J, **48, 70**
Marron, J, 61, 117
Matos, L, 59, 110
Mattos, T, 59, 110
McGarry, M, 53, 86
McKeague, I, **56, 97**
Mei, Y, **56, 99**
Meng, L, **49, 71**
Meng, Q, 62, 119
Mentch, L, 58, 108
Miao, G, 48, 68
Miao, H, **51, 81**
Molstad, A, **58, 106**
Mondal, S, 50, 75
Moran, G, **47, 65**
Morris, JC, 62, 119
Morris, N, 57, 100
Moulder, KL, 62, 119
Mueller, P, **57, 103**
Mueller-Velten, G, 50, 75
Mukherjee, R, 50, 78
Murphy, S, **59**
Murphy, SA, 54, 91
- Nahum-Shani, I, 54, 91
Needham, T, 51, 82
Nguyen, D, 59, 109
Ni, Y, 49, **60, 72, 113**
Ning, J, 53, 88
Niu, Y, **55, 93**
- O'malley, AJ, 60, 112
Olivari, R, 56, 98
Onnela, J, 50, 52, 78, 84
Ordonez, J, **56, 98**
- Pajewski, N, 54, 91
Pan, W, 55, 56, 96
Pan, Y, 58, 106
Park, Y, **54, 91**
Passoneau, R, 62, 120
Paulon, G, 57, 103
Pearl, D, 62, 120
Peiris, B, **61, 118**
Peng, Y, 49, 73
Perera, R, 55, 94
Perzanowski, M, **55, 94**
Peterson, E, **60, 114**
Peterson, K, 49, 73
Pitchford, A, **49, 73**
Platt, R, 48, 69
Podkopaev, S, **57, 100**
Prates, M, 56, 98
Pratola, M, **60, 114**
Price, B, **61, 117**
Prothero, J, **61, 117**
- Qi, J, **51, 81**
Qi, Y, 61, 116
Qian, W, 54, 91
Qin, F, 58, 105
Qin, L, **49, 72**
Qin, Q, **54, 92**
Qiu, P, **49, 49, 55, 73, 73, 74, 95**
Qiu, X, 61, 115
Qu, Y, **54, 90**
- Rahimi-Eichi, H, 52, 84
Ramdas, A, 57, 100
Rathouz, P, 48, 70
Reich, BJ, 50, 78
Ren, Z, 51, 52, 55, 81, 84, 95
Retnam, R, **60, 112**
Rhee, C, 59, 109
Rivera, A, 61, 115
Robison, L, 52, 85
Roeder, K, 51, 81
Rosen, O, 59, 109
Rosenblum, M, 62, 119
Rosenburg, P, 57, 104
Rosner, B, 51, 80
Roy, A, **49, 49, 71, 71**
Ruan, P, **57, 102**
Rubin, L, 49, 72
- Sachdeva, A, **61, 116**
Sahoo, I, **50, 78**
Samworth, R, 47, 66
Sang, Y, **57, 102**
Sarkar, A, 57, 103
Satten, G, 61, 115
Sayour, E, 57, 101
Scheffler, A, 60, 112
Schildcrout, J, **48, 48, 70, 70**
Schmidli, H, 50, 75
Schmitzer, M, 48, 69
Senturk, D, 59, 109
Sergin, N, 56, 99
Sevilimedu, V, **52, 83**
Sha, Y, 49, 72
Shan, G, **61, 119**
Shang, HL, **59, 110**
She, X, 49, 73
Shen, P, **49, 73**
Shen, S, **55, 95**
Shen, X, 55, 56, 96
Shentu, Y, 50, 75
Sheppard, L, 58, 107
Sherwood, B, **61, 61, 117, 117**
Shetty, V, 54, 91
Shih, M, 48, 69
Shojaie, A, 58, 107
Shortreed, S, 58, 107
Shu, D, **48, 51, 69, 80**
Siegmond, D, **47**
Signorovitch, J, 55, 94
Simon, N, 58, 107
Small, D, 50, 77

Smalley, K, 58, 104
 Song, P, 49, 50, 55, 61, 73, 77, 95, 116
 Song, R, 52, 84
 Song, X, **56, 98**
 Sridhar, D, 47, 65
 Sriram, TN, 57, 103
 Sriramareddy, S, 58, 104
 Srivastava, A, **51, 82**
 Srivastava, S, 60, 112
 Starr, J, 50, 78
 Stern, Y, 60, 112
 Su, Z, 54, **56, 90, 91, 97**
 Sun, (J), **48, 68**
 Sun, L, 60, 113
 Sun, W, 56, 98
 Sun, WW, **58, 106**
 Sun, Y, 50, **53–55, 75, 86, 93**
 Szpiro, A, 58, 107

 Tan, M, 57, 101
 Tang, W, **47, 65**
 Tang, X, **59, 111**
 Tao, R, 48, 70
 Taub, M, 62, 119
 Tian, H, 50, 75
 Tian, X, 53, 86
 Ting, N, **52, 60, 83, 113**
 Toh, S, 48, 51, 69, 80
 Tong, G, 54, 90
 Tong, W, **47, 67**
 Tran-Dinh, Q, 61, 117
 Trella, AL, **54, 91**
 Troendle, J, **53, 87**
 Tsao, P, **48, 69**
 Tseng, G, 51, 80, 81
 Tsung, F, **58, 105**
 Tu, W, 48, 68
 Tuft, M, 59, 109
 Tyx, R, 61, 115

 Valeri, L, **50, 52, 78, 84**
 Vemuri, B, **52, 82**
 Venkataraman, A, **61, 115**

 Viles, W, 60, 111
 Volgushev, S, 51, 80

 Waller, L, 60, 114
 Wang, C, **59, 111**
 Wang, G, **48, 50, 51, 62, 69, 76, 79, 119**
 Wang, H, 48, **53, 69, 89**
 Wang, J, **53, 55, 56, 88, 96, 97**
 Wang, L, **50, 51, 54, 60, 75, 80, 91, 114**
 Wang, M, 50, 76, 77
 Wang, P, **53, 88**
 Wang, R, 48, 51, 69, 80
 Wang, S, 48, **57, 59, 68, 101, 109, 109**
 Wang, T, 47, 66
 Wang, X, 50, 51, **53, 54, 58, 78, 79, 88, 93, 104**
 Wang, Y, 47, **49, 55, 60, 65, 74, 94, 112**
 Wang, Z, 62, 120
 Wegkamp, M, 47, 65
 Wei, L, **59**
 Wherry, EJ, 49, 72
 Williamson, B, **58, 107**
 Wright, C, **62, 119**
 Wu, C, 53, **55, 55, 86, 96, 96**
 Wu, CO, **53, 86**
 Wu, E, 55, 94
 Wu, F, 48, 69
 Wu, H, **51, 81**
 Wu, M, 61, 116
 Wu, Q, **52, 86**
 Wu, R, **58, 106**
 Wu, S, 48, 48, **68, 70**
 Wu, Z, **60, 61, 111, 118**

 Xi, D, 54, **60, 89, 113**
 Xia, Y, 51, 79
 Xiang, D, **58, 105**
 Xiao, F, 58, 105
 Xiao, H, 55, 93

 Xiao, J, 59, 111
 Xiao, Q, 53, 89
 Xie, D, 56, 97
 Xie, F, 60, 114
 Xie, X, **55, 95**
 Xiong, C, 62, 119
 Xiong, M, 48, 67
 Xiong, Y, **48, 70**
 Xu, G, 52, **59, 59, 85, 110, 111**
 Xu, M, 54, 92
 Xu, Q, 50, **58, 77, 105**
 Xu, T, **48, 67**
 Xu, Y, **49, 72**
 Xu, Z, 50, **57, 75, 101**
 Xue, H, 55, **56, 96, 96**

 Yan, H, 56, 99
 Yang, J, 53, 89
 Yang, K, **49, 74**
 Yang, M, 50, 75
 Yang, S, 53, 61, 88, 118
 Yang, X, 56, 100
 Yang, Y, 48, **49, 59, 69, 72, 110**
 Yao, F, **59, 109**
 Yao, S, 58, 104
 Ye, C, **61, 116**
 Ye, T, **52, 84**
 Ye, Z, **53, 88**
 Yi, G, **55, 93**
 Yin, A, 57, 101
 Yin, Q, 52, 83
 You, L, **49, 73**
 Young, JG, 48, 69
 Yu, L, 52, 83
 Yu, X, 52, **58, 85, 104**
 Yuan, A, **57, 101**
 Yung, G, 50, 75

 Zeng, D, 55, 94
 Zhan, T, **60, 113**
 Zhang, B, 49, 72
 Zhang, H, 61, 118
 Zhang, K, 53, **61, 89, 118**

 Zhang, KW, 54, 91
 Zhang, L, **47, 51, 58, 66, 79, 108**
 Zhang, P, **50, 75**
 Zhang, Q, 50, **53, 53, 76, 88, 89**
 Zhang, R, 58, 105
 Zhang, T, **57, 103**
 Zhang, W, **56, 59, 99, 108**
 Zhang, X, **51, 56, 58, 60, 80, 97, 106, 111**
 Zhang, Y, 48, 68
 Zhang, Z, **47, 65**
 Zhao, A, **50, 77**
 Zhao, H, 57, 102
 Zhao, J, 48, **60, 67, 113**
 Zhao, L, **50, 75**
 Zhao, N, 61, 115
 Zhao, Y, 57, 101
 Zhao, Z, **51, 61, 79, 118**
 Zheng, Q, **50, 77**
 Zhong, J, **55, 94**
 Zhong, K, **56, 98**
 Zhong, W, **57, 101**
 Zhou, B, **57, 102**
 Zhou, H, **47, 48, 65, 69**
 Zhou, J, **48, 48, 58, 60, 68, 69, 107, 111**
 Zhou, Q, 55, 93
 Zhou, S, **58, 108**
 Zhou, W, 61, 118
 Zhou, Y, **50, 55, 55, 77, 94, 95**
 Zhou, Z, **51, 55, 80, 95**
 Zhu, J, 47, 52, 65, 85
 Zhu, L, 52, 85
 Zhu, R, 56, 59, 97, 111
 Zhu, X, **61, 118**
 Zhu, Z, **47, 66**
 Zhuang, Y, 58, 107
 Zou, K, 49, 74
 Zou, S, 58, 105

**See you in 2023 ICSA
Applied Statistics Symposium
at the University of Michigan
in Ann Arbor, Michigan**

www.icsa.org



International Chinese Statistical Association

泛華統計協會