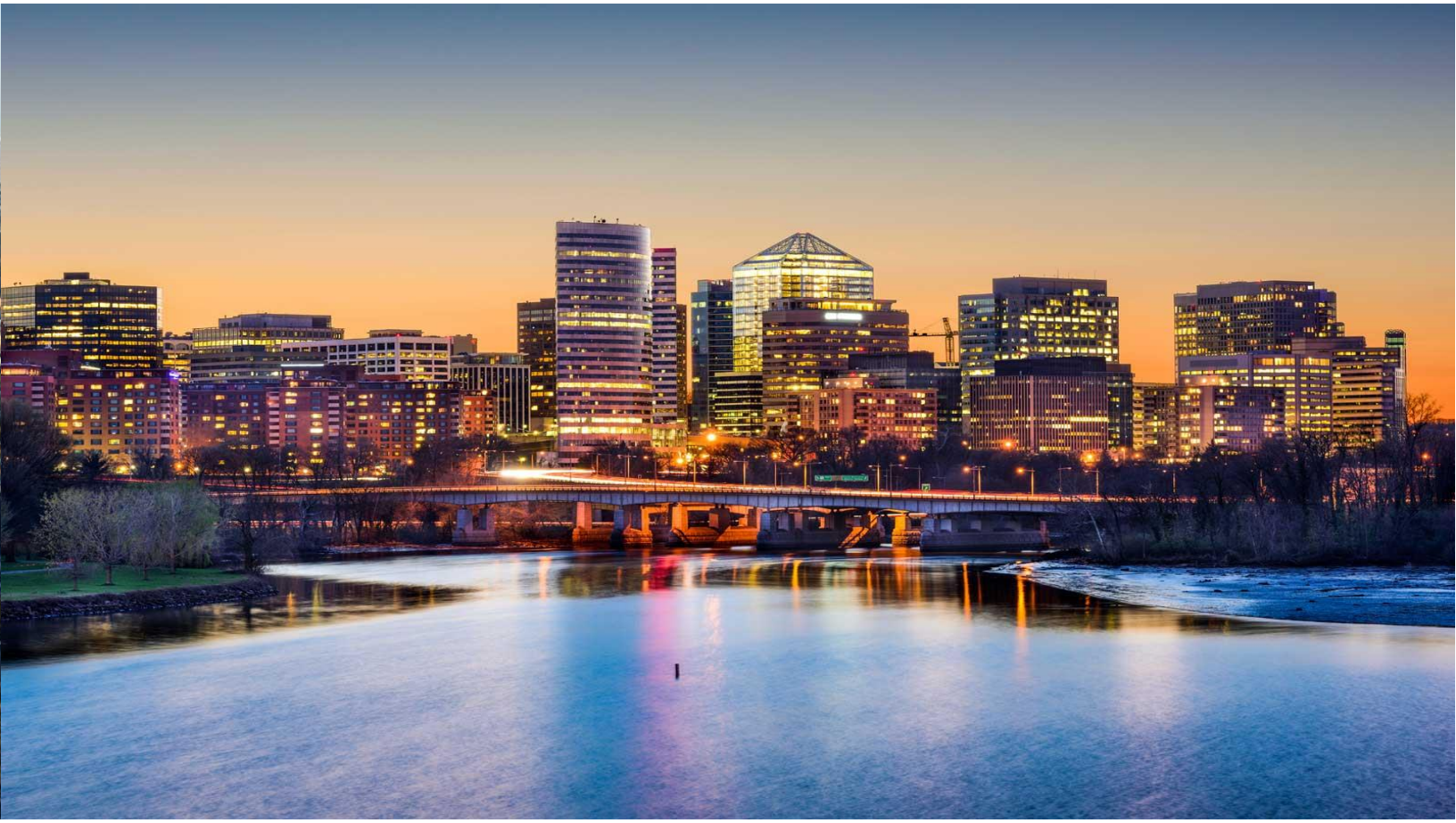

2021 ICSA APPLIED STATISTICS SYMPOSIUM

September 12-15

International Chinese Statistical Association

泛華統計協會

International Chinese Statistical Association

Applied Statistics Symposium

2021

CONFERENCE INFORMATION, PROGRAM AND ABSTRACTS

September 12 - 15, 2021

Organized by

International Chinese Statistical Association

Whova Login Instruction

To participate in the ICSA 2021 Applied Statistics Symposium, please sign up at https://whova.com/portal/webapp/virtu3_202109/ or download the Whova App. See the flyer below

Get Whova for International Chinese Statistical Association
2021 Applied Statistics Symposium

Official Even App

- Explore the **professional profiles** of event speakers and attendees
- Send **in-app messages** and **exchange contact info**
- **Network and find attendees** with common affiliations, educations, shared networks, and social profiles
- Receive **update notifications** from organizers
- Access the **even agenda**, GPS guidance, maps, and parking directions at your fingertips

Download Whova and take your event mobile.

Get Whova from the App Store or Google Play.

Please sign up for the app with your **social media account** or **email**

The event invitation code is:

e9wb0zxg2w

You will be asked for an event invitation code after installing Whova

Contents

Welcome Letter
The 2021 ICSA Applied Statistics Symposium

Dates: September 12-15, 2021
Format: Virtual via the *Whova* e-Platform

Welcome to the 2021 International Chinese Statistical Association (ICSA) Applied Statistics Symposium held virtually! This is the 30th annual symposium for ICSA.

It has been over one year and a half since the beginning of the COVID-19 pandemic, which has had an immense global impact and greatly shaped our lives both personally and professionally. As you may recall, the ICSA 2020 symposium was also held virtually last December. The ICSA 2021 symposium's Organizing Committee had hoped for an in-person symposium during the summer of 2021. However, at the last minute before signing a contract with Marriott, the Organizing Committee decided to continue considering a virtual format to protect the health and wellbeing of our participants. Thus, ICSA 2021 will be held on the e-platform *Whova*, September 12-15, 2021.

The Organizing Committee has developed a comprehensive scientific program that covers a broad spectrum of topics across multiple disciplines, including Statistics, Data Science, Computer Science, Biomedical Research, and other related fields. Themed **"Leading with Statistics and Innovation"**, ICSA 2021 features three distinguished Keynote Speakers from both Academia and Industry, including Dr. Scott Evans (George Washington University), Dr. Nicholas P. Jewell (London School of Hygiene and Tropical Medicine), and Dr. Ram Tiwari (Bristol Myers Squibb) throughout the entire conference. In addition, on the evening of September 14, 2021, Dr. Bin Yu (University of California, Berkeley) will give a live event speech on *veridical data science*, followed by the General Member Meeting, Awards Ceremony, and Social Activities (organized by Dr. Kelly H. Zou) consisting of a "Famous Landmarks Photo Contest" and a live Statistical Trivia game. The symposium also holds two special panels, entitled "Leadership and Communication for Statisticians and Data Scientists" (organized by Drs. Jiayang Sun and Hulin Wu) and "Statistics and Data Science Partnerships and Collaborations across Sectors" (organized by Dr. Kelly H. Zou).

Sponsored by 16 industry partners, ICSA 2021 offers 6 short courses, 108 invited sessions, a poster session, and a student paper awards session, as well as ample opportunities for networking and recruiting. The symposium attracts more than 600 participants worldwide. Despite being apart physically, we strive to bring people together to share innovative methods and cutting-edge applications, interact with fellow attendees, and positively impact our community, especially during the pandemic era.

Thank you for submitting your research and engaging with us before and during the ICSA 2021 Applied Statistics Symposium. We are confident that all of you find this virtual symposium stimulating, rewarding, and meaningful!

We look forward to welcoming you all to ICSA virtually and interactively!

Guoqing Diao, Ph.D., on behalf of the ICSA's 2021 Applied Statistics Symposium Executive and Organizing Committees (Primarily Based in Washington DC, USA)

EXECUTIVES

President: Colin Wu (wuc@nhlbi.nih.gov)
Past-President: Jianguo (Tony) Sun (sunj@missouri.edu)
President-Elect: Zhezhen Jin (zj7@columbia.edu)
Executive Director: Mengling Liu (2020-2022, executive.director@icsa.org)
ICSA Treasurer: Rochelle Fu (2019-2021, fur@ohsu.edu)

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

BOARD of DIRECTORS

Yinglei Lai (2019-2021, ylai@gwu.edu)
Lei Shen (2019-2021, shen_lei@lilly.com)
Yifei Sun (2019-2021, ys3072@cumc.columbia.edu)
Xin Tian (2019-2021, tianx@nhlbi.nih.gov)
Kelly Zou (2019-2021, KelZouDS@gmail.com)
Jason Liao (2020-2022, jason_liao@merck.com)
Bin Nan (2020-2022, nanb@uci.edu)
Peihua Qiu (2020-2022, pqiu@phhp.ufl.edu)
Jane Zhang (2020-2022, Zhang_Jane@Allergan.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)

COMMITTEES:

Program Committee

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

Xuming He (2019-2021, JSM Representative 2020, xmhe@umich.edu)
Aiyi Liu (2020-2022, JSM Representative 2021, liua@mail.nih.gov)
Pei Wang (2021-2023, JSM Representative 2022, pei.wang@mssm.edu)
Guoqing Diao (2020-2022, ICSA Symposium 2021, gdiao@email.gwu.edu)
Samuel Wu (2021-2023, ICSA Symposium 2022, samwu@biostat.ufl.edu)
Hongzhe Lee (2021-2022, ICSA International Conference 2019, hongzhe@pennmedicine.upenn.edu)
Xin-Yuan Song (2021-2022, ICSA International Conference 2022, xysong@sta.cuhk.edu.hk)
Alan Y Chiang (2019-2021, achiang@celgene.com)
Bin Nan (2019-2021, nanb@uci.edu)
Ji Zhu (2019-2021, jizhu@umich.edu)
Qingning Zhou (2020-2022, qzhou8@uncc.edu)
Liang Zhu (2020-2022, Liang.Zhu@uth.tmc.edu)
Jie Chen (2020-2022, jiechen0713@gmail.com)

Awards Committee

Chair: Judy Wang (judywang@gwu.edu)
Members:

Jie Chen (2019-2021, jiechen0713@gmail.com)
Ning Hao (2019-2021, nhao@math.arizona.edu)
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)
Ji Zhu (2021-2023, jizhu@umich.edu)

Nominating and Election Committee

Chair: Ming Tan (Ming.Tan@georgetown.edu)
Members:

Bo Huang (2019-2021, Bo.Huang@pfizer.com)
Chunling Liu (2019-2021, catherine.chunling.liu@polyu.edu.hk)
Yiyuan She (2019-2021, yshe@stat.fsu.edu)
Jiayang Sun (2019-2021, jsun@case.edu)
Hailong Cheng (2020-2022, hailong.cheng@sunovion.com)
Bin Zhang (2020-2022, Bin.Zhang@cchmc.org)

Special Lecture Committee

Chair: Gang Li (2021, vli@ucla.edu)
Members:

Haiqun Lin (2019-2021, haiqun.lin@yale.edu)
Huazhen Lin (2019-2021, linhz@swufe.edu.cn)
Aiyi Liu (2021-2023, liua@mail.nih.gov)
Gang Li (2021-2023, GLi@its.jnj.com)

Publication Committee

Chair: Ming-Hui Chen (2021, ming-hui.chen@uconn.edu)
Members:

Hongzhe Li (Co-Editors of SIB, hongzhe@mail.med.upenn.edu)
Joan Hu (Co-Editors of SIB, joan_hu@sfu.ca)
Rong Chen (Co-Editors of Statistica Sinica, rongchen@stat.rutgers.edu)
Su-Yun Huang (Co-Editors of Statistica Sinica, syhuang@stat.sinica.edu.tw)
Xiaotong Shen (Co-Editors of Statistica Sinica, xshen@umn.edu)
Ding-Geng (Din) Chen (Editor of ICSA book series, dinchen@email.unc.edu)
Ming Wang (Editor for ICSA Bulletin, mwang@phs.psu.edu)
Mengling Liu (Executive Director of ICSA, executive.director@icsa.org)

Membership Committee

Chair: Bo Fu (2021, bo.fu@abbvie.com)
Co-Chair: Liuquan Sun (2021, slq@amt.ac.cn)

Members:

Niansheng Tang (2019-2021, nstang@ynu.edu.cn)
Lei Shen (2019-2021, shen_lei@lilly.com)
Shuwei Li (2020-2022, lishuwstat@163.com)
Yifei Sun (2021-2023, ys3072@cumc.columbia.edu)

IT Committee

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

Archive Committee

Chair: Xin Tian (2019-2021, tianx@nhlbi.nih.gov)
Members:

Xin (Henry) Zhang (2021-2023, henry@stat.fsu.edu)
Naitee Ting (2021-2023, naitee.ting@boehringer-ingelheim.com)

Finance Committee

Chair: Rochelle Fu (2019-2021, fur@ohsu.edu**)**
Members:

Hongliang Shi (hongliangshi15@gmail.com)
Rui Feng (ruifeng@pennmedicine.upenn.edu)
Xin He (2021 ICSA Applied Statistics Symposium treasurer, xinhe@umd.edu)

Financial Advisory Committee

Chair: Rui Feng (ruifeng@upenn.edu)
Members:

Hongliang Shi (hongliangshi15@gmail.com)
Nianjun Liu, (liunian@indiana.edu)
Xiangqin Cui (xiangqin.cui@emory.edu)
Fang Chen (FangK.Chen@sas.com)
Rochelle Fu (fur@ohsu.edu)

Lingzi Lu Award Committee (ASA/ICSA)

Chair: Chan, Ivan (2019-2021, ivan.chan@abbvie.com)
Members:

Jichun Xie (2018-2020, jichun.xie@duke.edu)
Shelly Hurwitz (2017-2022, hurwitz@hms.harvard.edu)
Laura J Meyerson (2020-2022, laurameyerson@msn.com)

ICSA Representative to JSM Program Committee

Aiyi Liu (2021, liua@mail.nih.gov)
Pei Wang (2022, pei.wang@mssm.edu)

AD HOC COMMITTEES

2021 Applied Statistics Symposium

Chair: Wei Sun (wsun@fredhutch.org)

2021 JSM Local Committee

Chair: Wei Sun (wsun@fredhutch.org)

2021 ICSA China Conference
Chair: Yingying Fan (fanyingy@usc.edu)

2022 Applied Statistics Symposium

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

CHAPTERS

ICSA-Canada Chapter
Liqun Wang (**Chair,** Liqun.Wang@umanitoba.ca)

ICSA-Midwest Chapter
Li Wang (**Chair,** li.wang1@abbvie.com)

ICSA-Taiwan Chapter
Chao A. Hsiung (**Chair,** hsiung@nhri.org.tw)

Conference Committees

Executive Committee:

- Executive Committee Chair: Guoqing Diao, George Washington University
- Scientific Program Co-Chair: Judy Wang, George Washington University
- Scientific Program Co-Chair: Qing Pan, George Washington University
- Poster Session Committee Chair: Fang Jin, George Washington University
- Program Book and Website Committee Chair: Scott Bruce, Texas A&M
- Local/Online Platform Committee Co-Chair: Alan Wu, BeiGene
- Local/Online Platform Committee Co-Chair: Zhen Chen, National Institutes of Health
- Treasurer: Xin He, University of Maryland, College Park
- Student Paper Competition Committee: Lu Mao, University of Wisconsin, Madison
- Short Course Committee Chair: Yan Ma, George Washington University
- Fund Raising Committee Chair: Li Wang, Abbvie
- Social Activity Committee Chair: Kelly H. Zou, Viatris
- Strategic Advisor: Hulin Wu, University of Texas Health Science Center at Houston
- Strategic Advisor: Aiyi Liu, National Institutes of Health

Scientific Program Committee:

- Co-Chair: Judy Wang, George Washington University
- Co-Chair: Qing Pan, George Washington University
- Pramita Bagchi, George Mason U
- Ionut Bebu, George Washington U
- Qixuan Chen,Columbia
- Weihua Guan,U of Minnesota
- Chen Hu, JHU
- Chiungyu Huang, UCSF
- Yi Huang, UMBC
- Tracy Ke, Harvard
- Daniel Li,BMS
- Guodong Li,U of Hong Kong
- Jialiang Li,National U of Singapore
- Yun Li,UNC
- Yun Li, Upenn
- Jianchang Lin,Takeda Pharmaceuticals
- Xuewen Lu,U of Calgary
- Yajun Mei,Georgia Tech
- Yang Ning, Cornell
- Wanli Qiao,George Mason U
- Annie Qu,UC Irvine
- Sundaram Rajeshwari, NIH/NICHD
- Martin Slawski, George Mason U
- Ying Sun,King Abdullah U of S & T

- Yanlin Tang,East China Normal U
- Tiejun TONG, Hong Kong Baptist
- Diane Uschner, GWU
- Jingshen Wang, UC Berkeley
- Li Wang, Iowa State U
- Sijian Wang, Rutgers
- Tongtong Wu,Rochester U
- Grace Yi,U of Western Ontario
- Anru Zhang,U Wisconsin
- Zhiwei Zhang, NCI
- Yunpeng Zhao, Arizona State U
-

Poster Session Committee:

- Chair: Fang Jin, George Washington University
- Xiaoke Zhang, George Washington University
- Zhengling Qi, George Washington University
- Lin Wang, George Washington University

Student Paper Competition Committee:

- Chair: Lu Mao, UW-Madison
- Peng Ding, UC Berkeley
- Ni Zhao, JHU
- Noorie Hyun, Medical College of Wisconsin
- Fei Gao, Fred Hutch
- Edward Kennedy, CMU
- Rui Duan, Harvard
- Yifei Sun, Columbia U
- Lu Tang, U Pitt
- Zhenke Wu, UMich
- Muxuan Liang, Fred Hutch
- Ye Ting, Penn
- Weijie Su, Penn

Short Course Committee:

- Chair: Yan Ma, George Washington University
- Shuo Chen, University of Maryland School of Medicine
- Pan Wu, Vertex Pharmaceuticals
- Jinbo Chen, University of Pennsylvania
- Yong Ma, FDA

Social Activity Committee

- Chair: Kelly H. Zou, Viatris
- Xiao (Robin) Jing, Lakeport Capital
- Shiyang Ma, Columbia University Mailman School of Public Health
- Jim Z. Li, Viatris
- Mina Riad, Lotus Clinical Research, LLC
- Franklin Sun, ClinChoice
- Qiuyi (Queenie) Wu, University of Rochester Medical Center
- Ching-Ray Yu, Pfizer

Volunteers

Name	School
Aditi Arunmozhi	George Washington University
Fengnan Deng	George Mason University
Lizhao Ge	George Washington Universisty
Aisha Ghazwani	George Washington University
Muzhe Guo	George Washington University
Ningbo Jian	UMBC
Peng Jin	NYU
Andi Li	George Washington Universisty
Yiwei Li	NYU
Xiaoran Jiang	George Mason University
Xiaoyang Ma	Georgetown University
Shiyu Shu	George Washington Universisty
Juntao Su	George Washington Universisty
Yakun Wang	George Mason University
Siqi Wei	George Mason University
Qiuyi Wu	University of Rochester
Zhou Yang	George Washington Universisty
Jiebing Yin	University of Virginia
Weibin Zhong	George Mason University & BMS
Yizhao Zhou	University of Pennsylvania

Sponsors and Career Services

The 2021 ICSA Applied Statistics Symposium Program Committees gratefully acknowledge the generous support of our sponsors.

Gold sponsors

Sponsors and Career Services

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.
www.linkedin.com/company/beigene
@BeiGeneGlobal
BeiGene

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.

Bristol Myers Squibb™

Visit **bms.com** to see how we're bringing a human touch to everything we do.

Creating Possible

For more than 30 years, Gilead has created breakthroughs once thought impossible for people living with life-threatening diseases. We are a leading biopharmaceutical company with a pioneering portfolio and ever-expanding pipeline of investigational medicines.

Our commitment goes well beyond science. We innovate with the goal of eliminating barriers and providing access to healthcare for people who need it most.

For more information, please visit www.gilead.com.

Gilead is proud to support the International Chinese Statistical Association 2021 Applied Statistics Symposium.

GILEAD
Creating Possible

There has never been a more exciting time to be a biostatistician at GSK

Our purpose is to help people do more, feel better and live longer through the discovery and development of new medicines and vaccines. What we do is life-changing and lifesaving; however, the complexity of human biology means even the most brilliant scientific teams are destined to fail much more than they succeed. As biostatisticians, we are at the vanguard of GSK's response to these challenges.

We are transforming R&D decision-making from a process based on expert gut feel and past-experience to one anchored in robust quantitative examination and statistical analysis. Our work enables scientists to make data-driven choices about which potential new medicines and vaccines are most and least promising – ultimately saving time and money and maximising success.

We want to attract and invest in talented and committed statisticians like you to grow an industry-leading team who put quantitative rocket boosters under every project we touch, improving the pace, scale and success with which our company serves our patients. Join us!

Learn more at GSKBiostatisticsCareers.com

Sponsors and Career Services

The Lotus Group LLC is the leading recruitment firm for biometrics professionals in the Life Sciences industry.

We are honored to be recognized as one of the top 50 fastest-growing women-owned businesses in 2020 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. We pride ourselves on building solid, long-lasting relationships and providing exceptional, personalized service.

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.

If you are a talented statistician, statistical programmer, or data management professional, we have exciting opportunities for you!

Step on to The Lotus Group bridge – connecting great candidates and companies – and start your journey to success! www.tlgcareers.com

Sponsors and Career Services

AMGEN FACT SHEET | 2021

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.

AstraZeneca

AstraZeneca is a global, science-led biopharmaceutical company that focuses on the discovery, development and commercialisation of prescription medicines. Our Purpose is to push the boundaries of science to deliver life-changing medicines. We believe the best way we can achieve our Purpose is to put science at the centre of everything we do. Science defines who we are. It is why we come to work every day and is part of our DNA. But this is only half of the story. We know that we do not have all the answers. We want to share ideas because we believe it results in better medicines. We want the way we work to be inclusive, open and collaborative. This approach runs through all that we do. We focus on three main therapy areas – Oncology, Cardiovascular & Metabolic Disease (CVMD) and Respiratory – and we are also selectively active in the areas of autoimmunity, neuroscience and infection. For more information, visit www.astrazeneca.com

Sponsors and Career Services

MERCK
INVENTING FOR LIFE

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.

Sanofi is a global life sciences company committed to creating health, social, economic and scientific value for society and transforming the practice of medicine for patients around the world. More than international 110,000 employees are dedicated to making a difference in patients' daily lives, transforming scientific innovation into healthcare solutions primarily focused oncology, immunology, rare disease and neurology.

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 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

Sponsors and Career Services

VERTEX
THE SCIENCE *of* POSSIBILITY

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

Sponsors and Career Services

At Johnson & Johnson, we believe good health is the foundation of vibrant lives, thriving communities and forward progress. That's why for more than 130 years, we have aimed to keep people well at every age and every stage of life. Today, as the world's largest and most broadly-based health care company, we are committed to using our reach and size for good. We strive to improve access and affordability, create healthier communities, and put a healthy mind, body and environment within reach of everyone, everywhere. We are blending our heart, science, and ingenuity to profoundly change the trajectory of health for humanity.

High Quality Drug & Medical Devices Development in the U.S and China CR Medicon is a fast-growing, innovative Contract Research Organization (CRO) dedicated to providing high-quality clinical development services globally. With close to 600 staff, we are a full-service platform in China with a dedicated biometrics service platform in U.S. Our team is led by industry veterans with 15-25 years of experience. We have adopted the Medidata technology platform globally and we are accredited in several Medidata products including Rave/Balance/CTMS/eTMF. Our team is highly proficient in CDISC implementation including CDISC CDASH/SDTM/ADaM. We can help implement the same set of global standards achieving top quality deliverables to meet the requirements of both US FDA and China NMPA. If you are thinking about expanding your drugs or devices into the China market, or if you are evaluating a biometrics CRO that can provide high quality services yet at the same time can always be flexible, CR Medicon is your best choice.

Data Management & Biometrics Services in the U.S. We have a large Data Management and Biostatistics team with hundreds of successful study and submission experiences in US and China. Our team is led by industry veterans with 15-20 years of experience. We use the Medidata system for our U.S. studies and about 75% of our China studies. The vast majority of our studies are implemented using CDISC standards. Our biometrics team is ready to help you to conduct studies meeting international data and reporting standards both in the U.S. and China.

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.

Landmarks Photo Contest

Join us on Sep. 14 for

ICSA 2021 "Brilliance!" landmarks photo contest

We will hold the meeting on Tuesday evening for

- Famous Landmarks Photo Contest Winners
- Photo contest top landmarks trivia – live
- Statistics Trivia game – live
- Special guest appearance associated with the Statistics Trivia

ICSA 2021 Applied Statistics Virtual Symposium Schedule

Activity/Event Schedules (All times are US Eastern Time)			
Time	**Location**	**Activity**	**Host**
9/12: Sunday	Virtual Front Desk Open from 8:00AM-8:00PM		
8:30-12:30PM	Breakout rooms	Morning short courses	Short Course Committee
12:30-1:30PM	Lunch break		
1:30-5:30PM	Breakout rooms	Afternoon short courses	Short Course Committee
8:30-5:30PM	Breakout rooms	Full-day courses	Short Course Committee
6:00-7:40PM	Breakout rooms	Parallel Sessions	Session Chairs
9/13: Monday	Virtual Front Desk Open from 8:00AM-10:00PM		
8:45-9:00AM	Virtual Main Room	Welcome and opening ceremony	Colin Wu & Guoqing Diao
9:00-10:00AM		Keynote Talk I: Scott Evans	Guoqing Diao
10:00-10:20AM	Coffee Break		
10:20-12:00PM	Breakout rooms	Parallel Sessions	Session Chairs
12:00-12:20PM	Lunch break:		
12:20-1:50PM	Virtual Main Room	Panel Discussion Session: Leadership and Communication for Statisticians and Data Scientists	Jiayang Sun
2:00-3:40PM	Breakout rooms	Parallel Sessions	Session Chairs
3:40-4:00PM	Coffee Break		
4:00-5:40PM	Breakout rooms	Parallel Sessions	Session Chairs
6:00-9:30PM	Virtual Poster & Mixer Room	Poster Sessions & Mixer	Poster Session Chairs
9/14: Tuesday	Virtual Front Desk Open from 8:30AM-7:30PM		
9:00-10:00AM	Virtual Main Room	Keynote Talk II: Ram Tiwari	Margaret Gamalo
10:00-10:20AM	Coffee Break		
10:20-12:00PM	Breakout rooms	Parallel Sessions	Session Chairs
12:00-12:20PM	Lunch break		
12:20-1:50PM	Virtual Main Room	Panel Discussion Session: Statistics and Data Science Partnerships and Collaborations across Sectors	Kelly Zou
2:00-3:40PM	Breakout rooms	Parallel Sessions	Session Chairs
3:40-4:00PM	Coffee Break		
4:00-5:40PM	Breakout rooms	Parallel Sessions	Session Chairs
7:30-10:00PM	ICSA Zoom Room	ICSA 2021 General Member Meeting, Awards Ceremony	Mengling Liu Colin Wu Guoqing Diao Kelly Zou
9/15: Wednesday	Virtual Front Desk Open from 8:30AM-10:30AM		
9:00-10:00AM	Virtual Main Room	Keynote Talk III: Nicholas P. Jewell	Colin Wu
10:00-10:20AM	Coffee Break		
10:20-12:00PM	Breakout rooms	Parallel Sessions	Session Chairs
12:00PM	Adjournment		

Keynote Speaker

Scott Evans, Ph.D., Professor and Founding Chair of the Department of Biostatistics Bioinformatics and the Director of the George Washington Biostatistics Center. He is the Director of the Statistical and Data Management Center (SDMC) for the Antibacterial Resistance Leadership Group (ARLG) and the Co-chair of the Benefit-Risk Balance for Medicinal Products Committee for the Council for International Organizations of Medical Sciences (CIOMS). Dr. Evans is a recipient of the Mosteller Statistician of the Year Award and the Founders Award from the American Statistical Association (ASA), the Robert Zackin Distinguished Collaborative Statistician Award, an elected member of the International Statistical Institute (ISI), and is a Fellow of the ASA, Society for Clinical Trials (SCT), and the Infectious Disease Society of America (IDSA).

Time: September 13 (Monday): 9:00-10:00AM (Eastern Time)
Host: Guoqing Diao, Ph.D., ICSA 2021 Applied Statistics Symposium Executive Committee Chair and Professor, Department of Biostatistics and Bioinformatics, George Washington University

Title: The Catalyst to Better Answers: More Thoughtful Questions

Abstract: We often debate how to design, conduct, or analyze a study. At its core, what we are debating is not the answer, but the question. Once the intricacies of the question are well defined and understood, the path forward becomes clear. Unfortunately, we often fail to recognize the most important questions and tailor the design, conduct, and analysis of studies to address these questions. As a result, we fail to get optimal answers to the most important questions.

For example, randomized clinical trials are the gold standard for evaluating the benefits and harms of interventions but often fail to provide the necessary evidence to inform practical medical decision-making. Typical analyses of clinical trials involve intervention comparisons for each efficacy and safety outcome. Outcome-specific effects are estimated and potentially combined in benefit:risk analyses with the belief that such analyses inform the totality of effects on patients. However summing marginal analyses of each outcome does not effectively characterize the effects on patients. Such approaches do not incorporate associations between outcomes of interest or the cumulative nature of component outcomes on patients, suffer from competing risk challenges when interpreting outcome-specific results, and since efficacy and safety analyses are conducted on different analysis populations, the population to which these analyses generalize, is unclear.

Identification of the most important clinical questions and adjusting our approaches to address these questions is the greatest opportunity to advance medicine and public health. In clinical trials and diagnostics studies, increased interest on questions of a pragmatic origin is needed to match their clinical importance and real-world utility. Ideas for adjustments to trial design, conduct, analyses, and reporting, to answer the most important questions for medical-decision making, are discussed.

Keynote Speaker

Nicholas P. Jewell, Ph.D., Chair in Biostatistics and Epidemiology at the London School of Hygiene and Tropical Medicine. Previously he was Professor of Biostatistics and Statistics at the University of California, Berkeley (1981-2018) where he was Vice Provost for six years. Jewell received his PhD in Mathematics from the University of Edinburgh, and was Assistant Professor of Statistics at Princeton University before moving to Berkeley. He has also held visiting appointments at Oxford University, the Karolinska Institutet, and the University of Kyoto. Jewell was elected to the National Academy of Medicine in 2017. He is a Fellow of the American Statistical Association, the Institute of Mathematical Statistics, and the American Association for the Advancement of Science. He received the 2005 Snedecor Award, the Harvard University 2012 Marvin Zelen Leadership Award in Statistical Science, the 2018 Cupples Award for Excellence in Teaching, Research, and Service in Biostatistics from Boston University, and the 2021 Nathan Mantel Award from the ASA for lifetime contributions to the development and application of statistical science to problems and issues in epidemiology. Jewell has published over 200 articles in statistics, mathematics, epidemiology, medicine, and history. He is the author of *Statistics in Epidemiology*, and *Causal Inference in Statistics: A Primer*, with Judea Pearl and Madelyn Glymour.

Time: September 15 (Wednesday): 9:00-10:00AM (Eastern Time)
Host: Colin Wu, Ph.D., ICSA President and Mathematical Statistician, National Heart, Lung, and Blood Institute, National Institutes of Health

Title: Test-negative designs: From Dengue and Ebola to COVID-19

Abstract: Test-negative designs are a relatively recent addition to observational study tool boxes, originally largely used for assessment of the seasonal influenza vaccination program. I will review the rationale and development of the original test-negative design and discuss extension to cover randomized exposures, and various kinds of clustering with illustrations from both dengue fever and ebola virus diseases. Recent interest has focused on uses of the test-negative design to estimate and compare effectiveness of COVID-19 vaccines. The presentation will highlight statistical issues associated with the design and analysis of resulting data.

Keynote Speaker

Ram C. Tiwari, Ph.D., Head of Statistical Methodology at BMS since February 1, 2021. His prior services include serving at FDA, NCI/NIH and in academia. He received his MS and PhD degrees from Florida State University in Mathematical Statistics. He is a Fellow of the *American Statistical Association* and a past President of the *International Indian Statistical Association*. Dr. Tiwari has over 200 publications on statistical methods, and a newly authored book on "Signal Detection for Medical Scientists: Likelihood Ratio Test-based Methodology" published by Francis &Taylor.

Time: September 14 (Tuesday): 9:00-10:00AM (Eastern Time)
Host: Margaret Gamalo, Ph.D., Senior Director - Biostatistics, Pfizer Innovative Health

Title: Leveraging External Evidence for Augmenting a Single-Arm Study

Abstract: There is a wide variety of study designs that involve the leveraging of external data. These external data can be leveraged to augment a single-arm study, or construct or augment either the control arm or the treatment arm (or both) of a comparative clinical study, and they may come from a single source or from multiple sources. In this talk, I will use illustrative examples to present novel propensity-score based methods for leveraging external data to augment a single-arm study or to augment the control arm of a randomized controlled trial (RCT). These methods can be utilized for medical drug/device development.

Live Event Speaker

Bin Yu is Chancellor's Distinguished Professor and Class of 1936 Second Chair in the departments of statistics and EECS at UC Berkeley. She leads the Yu Group which consists of 15-20 students and postdocs from Statistics and EECS. She was formally trained as a statistician, but her research extends beyond the realm of statistics. Together with her group, her work has leveraged new computational developments to solve important scientific problems by combining novel statistical machine learning approaches with the domain expertise of her many collaborators in neuroscience, genomics and precision medicine. She and her team develop relevant theory to understand random forests and deep learning for insight into and guidance for practice. Moreover, **she is finishing a book "Veridical data science" (MIT Press) with her postdoc Rebecca Barter.**

She is a member of the U.S. National Academy of Sciences and of the American Academy of Arts and Sciences. She is Past President of the Institute of Mathematical Statistics (IMS), Guggenheim Fellow, Tukey Memorial Lecturer of the Bernoulli Society, Rietz Lecturer of IMS, and a COPSS E. L. Scott prize winner. She holds an Honorary Doctorate from The University of Lausanne (UNIL), Faculty of Business and Economics, in Switzerland. She is serving on the editorial board of Proceedings of National Academy of Sciences (PNAS) and the scientific advisory committee of the UK Turing Institute for Data Science and AI.

Time: September 14 (Tuesday): 7:30-10:00PM (Eastern Time)

Title: Veridical data science: how to teach data science to positively impact the world?

ICSA 2021 Applied Statistics Symposium

Two Virtual Panels on Leadership, Communication, Collaboration and Partnership

From 12:20 PM to 1:50 PM, U.S. Eastern Time (ET), September 13 and 14, 2021

The 2021 ICSA Applied Statistics Symposium (https://symposium2021.icsa.org), held virtually from September 12 to 15, 2021, is the 30th annual symposium for the International Chinese Statistical Association (ICSA). The conference theme is *Leading with Statistics and Innovation*.

Panel 1: *Leadership and Communication for Statisticians and Data Scientists*

From 12:20 PM to 1:50 PM, ET, September 13, 2021

Panelists: Hulin Wu, University of Texas Health Science Center at Houston; Xihong Lin, Harvard University; Colin Wu, NIH; Sylva Colins, FDA; Ruixiao Lu, Dahshu; and Catherine Truxillo, SAS. Moderator: Jiayang Sun, George Mason University.

Paul J. Meyer said: "Communication – the human connection – is the key to personal and career success." Statisticians or Data Scientists (SDS), who communicate effectively, can be leaders in many ways. As the world changes and we prepare to return to a new normal after the pandemic, we must discuss issues for developing a whole person to meet new global challenges. This session brings in excellent leaders in academia, government, and industry to discuss their perspectives about Leadership and Communication for Statisticians and Society. Topics will include essential elements of effective leadership/communication, formal and informal training and mentoring, roles of IT or social media, network, ethics, culture, and psychological aspects, as well as tips specifically for women, Asians, and general SDS. The session will end with a Q&A open to participants.

Panel 2: *Statistics and Data Science Partnerships and Collaborations across Sectors*

From 12:20 PM to 1:50 PM, ET, September 14, 2021

Panelists: Victoria Gamerman, Boehringer-Ingelheim; John E. Kolassa, Rutgers, The State University of New Jersey; Jim Z. Li, Viatris; Fanni Natanegara, Eli Lilly and Company; Kimberly Sellers, Georgetown University; Aniketh Talwai, Medidata, a Dassault Systèmes company; Moderator: Kelly H. Zou, Viatris.

Collaborations and partnerships can come in all shapes and forms. Martin Luther King, Jr.'s words may resonate: "We may have all come on different ships, but we're in the same boat now." Frequently, sharing of ideas between stakeholders from different organizations leads to exchange visits, support for graduate students, consulting jobs, grant support, and continuing education opportunities for statisticians or data scientists outside of academe. Although these activities statistical and data science problems from outside academia become use cases studies. The intellectual exchange that results is a key component of such partnerships. This panel includes experts from industry and consulting, which will have a wide appeal, given the increasing focus on inter-disciplinary research and the emergence of complex and high dimensional data. In particular, such challenges are common in health care research. In this invited session, several panelists discuss the key elements to form and sustain successful collaborations and partnerships, along with challenges and barriers. The panel discussion can be valuable to statisticians and data scientists in diverse areas and sectors.

If you have any questions, please send an email to symposium2021@icsa.org.

ICSA Applied Statistics Symposium Student Paper Awards

- Subhankar Bhadra, NC State University
  Title: Scalable community detection in massive networks via predictive inference
  See session 1 in the program

- Ying Zhou, University of Toronto
  Title: The Promises of Parallel Outcomes
  See session 1 in the program

- Rui Miao, The George Washington University
  Title: A Wavelet-Based Independence Test for Functional Data with an Application to MEG Functional Connectivity
  See session 1 in the program

- Ziyan Yin, Temple University
  Title: Linear Models with Distributed Data and Communication-Efficient Estimator: Limiting Distribution and Its Approximation
  See session 1 in the program

- Yujia Deng, University of Illinois, Urbana-Champaign
  Title: Query-augmented Active Metric Learning
  See session 1 in the program

- Bing Li, Brown University
  Title: Transportation of Area Under the ROC curve to a Target Population
  See session 1 in the program

Jiann-Ping Hsu Pharmaceutical and Regulatory Sciences Student Paper Award

- Manyun Liu, Georgia Southern University
  Title: Evaluation of SIMEX extrapolation methods in Accelerated failure time models with covariate measurement error
  See session 89 in the program

- Stephan Ogenstad, Jiann-Ping Hsu College of Public Health, Georgia Southern University
  Title: Principles of leading with statistical knowledge and a problem-solving approach to innovation
  See session 89 in the program

- Kao-Tai Tsai, BMS
  Title: Statistical Data Analysis in Pharmaceutical Industry
  See session 89 in the program

SC01: Large-Scale Spatial Data Science

Length: Half-Day

Instructors: Dr.Marc Genton (King Abdullah University of Science and Technology); Dr.Huang Huang (King Abdullah University of Science and Technology); Dr. Sameh Abdulah (King Abdullah University of Science and Technology)

Outline/Description: Spatial data science is concerned with analyzing the spatial distributions, patterns, and relationships of data over a predefined geographical region. It relies on the dependence of observations where the primary assumption is that nearby spatial values are associated in a certain way. For decades, the size of most spatial datasets was modest enough to be handled by exact inference. Nowadays, with the explosive increase of data volumes, High-Performance Computing (HPC) has become a popular tool for many spatial applications to handle massive datasets. Big data processing becomes feasible with the availability of parallel processing hardware systems such as shared and distributed memory, multiprocessors, and GPU accelerators. In spatial statistics, parallel and distributed computing can alleviate the computational and memory restrictions in large-scale Gaussian random process inference. In this course, we will first briefly cover the motivation, history, and recent developments of statistical methods so that the participants can have a general overview of spatial statistics. Then, the cutting-edge HPC techniques and their application in solving large-scale spatial problems with the new software ExaGeoStat will be presented.

About the Instructors: Marc G. Genton is a Distinguished Professor of Statistics at the Spatio-Temporal Statistics and Data Science (STSDS) research group, King Abdullah University of Science and Technology (KAUST), Saudi Arabia. He received his Ph.D. in Statistics in 1996 from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. He also holds an M.S. degree in Applied Mathematics teaching from EPFL. Prior to joining KAUST, he held faculty positions at MIT, North Carolina State University, the University of Geneva, and Texas A&M University. Most of Prof. Genton's research centers around spatial and spatio-temporal statistics. His interests include the statistical analysis, visualization, modeling, prediction, and uncertainty quantification of spatio-temporal data, with applications in environmental and climate science, renewable energies, geophysics, and marine science. He is a Fellow of the American Statistical Association, of the Institute of Mathematical Statistics, of the American Association for the Advancement of Science, and an elected member of the International Statistical Institute. In 2010, he received the El-Shaarawi award for excellence from the International Environmetrics Society (TIES) and the Distinguished Achievement Award from the Section on Statistics and the Environment (ENVR) of the American Statistical Association (ASA). In 2017, he received the Wilcoxon Award for Best Applications Paper in the journal Technometrics. He is the 2020 Georges Matheron Lecturer of the International Association for Mathematical Geosciences. Personal webpage: http://stsds.kaust.edu.sa Huang Huang is a research scientist at the Spatio-Temporal Statistics and Data Science (STSDS) research group, King Abdullah University of Science and Technology (KAUST), Saudi Arabia. Before working at KAUST, Huang was a postdoctoral fellow at the National Center for Atmospheric Research (NCAR), the Statistical and Applied Mathematical Sciences Institute (SAMSI), and Duke University, focusing on spatio-temporal statistical inference for large climate data sets. He received his Ph.D. in Statistics in 2017 from KAUST and M.S. and B.S. in Mathematics in 2014 and 2011 from Fudan University, China. His research interests include spatio-temporal statistics, functional data analysis, Bayesian modeling, machine learning, and High-Performance Computing (HPC) for large climate data. Personal webpage: https://hhuang90.github.io Sameh Abdulah is a research scientist at the Extreme Computing Research Center (ECRC), King Abdullah University of Science and Technology, Saudi Arabia. Sameh was a postdoctoral fellow at ECRC from 2016 to 2019. He received his M.S. and Ph.D. degrees from the Ohio State University, Columbus, USA, in 2014 and 2016. His work is centered around High-Performance Computing (HPC) applications, bitmap indexing in big data, large spatial datasets, parallel statistical applications, algorithm-based fault tolerance, and machine learning and data mining algorithms. Sameh is also a reviewer in several HPC-related journals and conferences. Personal webpage: https://sites.google.com/view/samehabdulah

SC02: Statistical methods for composite time-to-event outcomes

Length: Half-Day

Instructors: Dr. Lu Mao (University of Wisconsin-Madison)

Outline/Description: This course provides an overview of the recent statistical methodology developed for the analysis of composite time-to-event outcomes. Such outcomes typically combine death and (possibly recurrent) nonfatal events, such as hospitalization, tumor progression, or infection, and are routinely used as the primary efficacy endpoint in modern phase-III clinical trials. The traditional approach to composite outcomes is to focus on time to the first event, whichever type it is. Recent years have seen a surge of novel statistical methods, attracting the attention of both methodologists and practitioners. These include the win ratio (Pocock et al., 2012) and its various extensions, the restricted mean time in favor of treatment (an extension of the restricted mean survival time), generalized semiparametric proportional odds regression models, and so on. The new methods improve upon the existing ones in: 1. Proper prioritization of death over nonfatal events; 2. Fuller utilization of multiple/recurrent events; 3. Clear and interpretable definition of estimands in compliance with the ICH E9 (R1) guideline; 4. Versatile modeling strategies for general outcome types In addi-

tion, a number of user-friendly R-packages that implement the new methods have also become available. This course will supplement methodological studies with hands-on analysis of real-world data using R.

About the Instructors: Dr. Lu Mao joined the Department of Biostatistics and Medical Informatics (BMI) at UW-Madison as an Assistant Professor after finishing his doctoral dissertation on regressions of competing risks data at UNC Chapel Hill in 2016. His current research interests include survival analysis, causal inference, semiparametric theory, and clinical trials. He currently serves as the PI of an NIH R01 grant on statistical methodology for composite time-to-event outcomes in cardiovascular clinical trials and an NSF grant on causal inference in randomized trials with noncompliance. Besides methodological research, he also engages in collaborative studies in cardiology, radiology, cancer, and health behavioral interventions, where time-to-event and longitudinal data are routinely collected.

SC03: Biomarker discovery and pathway enrichment analysis of omics data

Length: Half-day
Instructor: Dr. Ali Rahnavard (George Washington University); Dr.Himel Mallick (Merck Research Laboratories, Merck & Co., Inc)

Outline/Description: Methodological advancements paired with measured multiomics data using high-throughput technologies enable capturing comprehensive snapshots of biological activities. In particular, low-cost, culture-independent omics profiling has made metagenomics, metabolomics, and proteomics ("multiomics") surveys of human health, other hosts, and the environment. The resulting data have stimulated the development of new statistical and computational approaches to analyze and integrate omics data, including human gene expression, microbial gene products, metabolites, and proteins, among others. Multiomics data generated from diverse platforms are often fed into generic downstream analysis software without proper appreciation of the inherent data differences, resulting in incorrect interpretations. Further, there are also an extensive collection of downstream analysis software platforms, and appropriately selecting the best tool can be extraneous for untrained researchers. This workshop will thus present a high-level introduction to computational multiomics, highlighting the state-of-the-art in the field and outstanding challenges geared towards downstream analysis methods. The workshop will include introducing typical multiomics studies' biological goals and the statistical methods currently available to achieve them.

SC04: Mining Electronic Health Record (EHR) data: the past, presence, and future

Length: Half-Day
Instructor: Dr. Shuangge Ma (Yale School of Public Health)

Outline/Description: With the fast development of information technologies and computing power, many nations (both central and local governments), large insurance systems, and health care systems have established or are establishing large electronic health record (EHR) databases. Effectively mining such databases using advanced data science techniques can generate unprecedented value for public health care, biomedicine, insurance management and planning, and other purposes. In this course, we will review the past and current status of the establishment of EHR databases and discuss successful (and unsuccessful) examples. Simple and advanced data science techniques that have been developed tailored to EHR data will be reviewed, with extensive examples provided. Discussions will then be provided on the future of EHR data mining, what needs to be done in terms of database/system development and data science methodology development, their implications, and the unique role statisticians can play.

About the Instructor: Shuangge Ma is Professor of Biostatistics at Yale University. His research interests include big data, EHR (electronic health record) data analysis, network analysis, public health, and health economics. He is an Elected Member of ISI and Fellow of ASA. He has published over 200 articles in prestigious journals and 2 books. He has been playing a leading role in biostatistical studies on cancer, health economics, and other areas. He has taught multiple short courses on advanced data science technologies at international conferences and renowned universities. He is an associate editor of JASA and other seven prestigious journals.

SC05: Functional Data Analysis Using R

Length: One-Day
Instructors: Dr. Jiguo Cao (Simon Fraser University)

Outline/Description: Functional data analysis (FDA) is a growing statistical field for analyzing longitudinal trajectories, curves, images or any manifold objects. FDA treats each random function observed over the whole domain as a sample element. Functional data can be commonly be found in many applications such as fitness data from the wearable device, air pollution, longitudinal studies, time-course gene expressions and brain images. This short course will cover the major FDA methods such as functional principal component analysis and functional linear regression models. All these methods will be demonstrated with real data applications using the R programming language. The goal of this short course is that trainees can learn how to use and develop FDA methods to analyze data.

About the Instructors: Dr. Jiguo Cao is the Canada Research Chair in Data Science and Professor at the Department of Statistics and Actuarial Science, Simon Fraser Univer-

sity. He is a prolific researcher who addresses critical, data-driven applications with well-tuned new models and estimation methods. He is a dedicated scholar with a worldwide reputation, a collaborator with a rich array of scientists, and an outstanding mentor of future scholars. Over a short period of time, Dr. Cao has made inroads across an impressively diverse range of sub-disciplines of statistics and data science; he builds on the synergies between functional data analysis (FDA), differential equations and big data to unveil novel statistical methods that greatly enhance users' understanding of their data. His high-impact, statistical frameworks are applied to real-world problems across various disciplines, including neuroscience, pharmacology, genetics, ecology, environment, and engineering. Dr. Cao's research is highly visible, with over 70 refereed publications and the delivery of 9 prestigious short courses on FDA and 74 invited talks and seminars in Australia, Canada, China, and USA.

SC06: Statistical and algorithmic foundations of reinforcement learning

Length: One Day
Instructors: Dr. Yuting Wei (Carnegie Mellon University); Dr. Yuejie Chi (Carnegie Mellon University); Dr. Yuxin Chen (Princeton University); Dr. Zhengyuan Zhou (New York University)

Outline/Description: Reinforcement learning (RL) studies how an agent learns to make sequential decisions to improve its performance over time through its interactions with an unknown environment. In the past few years, rapid advances in algorithms and explosive growth of computational and memory resources have enabled tremendous empirical success. In this short course, we convey this well-founded excitement by presenting a modern tutorial on the foundation of RL, both as a field and as a set of ideas. Our tutorial underscores important statistical and algorithmic developments in RL, both new and classic. We start by presenting a unified framework of the RL problem—from multi-arm bandits to general Markov decision processes—and then introduce classical planning algorithms when precise descriptions of the environments are available (Part I). Equipped with this background, we discuss two prominent approaches—model-based and model-free RL—assuming access to a generative model of the environment (Part II). Next, we discuss how to design intelligent exploration to optimize the agent's long-term performance when adaptively acting in a real environment (Part III). Finally, we present another distinctive approach: policy optimization, and conclude with further pointers to the literature (Part IV). Through this intense course, we hope to bring students and researchers in the statistical disciplines quickly up to speed to the heart of important insights that underlie the modern RL success.

About the Instructors: Dr. Yuting Wei is currently an Assistant Professor in the Statistics and Data Science Department at Carnegie Mellon University. Prior to that, she was a Stein Fellow at Stanford University, and she received her Ph.D. in statistics at University of California, Berkeley working with Martin Wainwright and Aditya Guntuboyina. She was the recipient of the 2018 Erich L. Lehmann Citation from the Berkeley statistics department for her Ph.D. dissertation in theoretical statistics. Her research interests include high-dimensional and non-parametric statistics, statistical machine learning, and reinforcement learning. Dr. Yuejie Chi is an Associate Professor in the department of Electrical and Computer Engineering, and a faculty affiliate with the Machine Learning department and CyLab at Carnegie Mellon University, where she held the Robert E. Doherty Early Career Development Professorship from 2018 to 2020. She received her Ph.D. and M.A. from Princeton University, and B. Eng. (Hon.) from Tsinghua University, all in Electrical Engineering. Her research interests lie in the theoretical and algorithmic foundations of data science, signal processing, machine learning and inverse problems, with applications in sensing systems, broadly defined. Among others, Dr. Chi received the Presidential Early Career Award for Scientists and Engineers (PECASE), the inaugural IEEE Signal Processing Society Early Career Technical Achievement Award for contributions to high-dimensional structured signal processing, and was named a Goldsmith Lecturer by IEEE Information Theory Society. Dr. Yuxin Chen is currently an assistant professor in the Department of Electrical and Computer Engineering at Princeton University, and is affiliated with Applied and Computational Mathematics, Computer Science, and Center for Statistics and Machine Learning. Prior to joining Princeton, he was a postdoctoral scholar in the Department of Statistics at Stanford University, and he completed his Ph.D. in Electrical Engineering at Stanford University. His research interests include high-dimensional statistics, mathematical optimization, and reinforcement learning. He has received the Princeton graduate mentoring award, the AFOSR Young Investigator Award, the ARO Young Investigator Award, the ICCM best paper award (gold medal), and was selected as a finalist for the Best Paper Prize for Young Researchers in Continuous Optimization. Dr. Zhengyuan Zhou is an assistant professor at the Stern School of Business, New York University. Before joining NYU Stern, he spent the year 2019-2020 as a Goldstine research fellow at IBM research. He received his BA in Mathematics and BS in Electrical Engineering and Computer Sciences, both from UC Berkeley. Subsequently, he has received a Master's in Computer Science, a Master's in Statistics, a Master's in Economics and a PhD in Electrical Engineering (with minors in Mathematics and Management Science & Engineering), all from Stanford University in 2019. His research interests include contextual bandits, reinforcement learning and data-driven sequential decision making problems at large.

Scientific Program (Sep. 12-15)

Sep. 12 6-8:05pm (EDT)

Session 1: Student Paper Competition Winners

Organizer: Lu Mao.
Chair: Lu Mao.

18:00(EDT) Query-augmented Active Metric Learning
♦*Yujia Deng*[1]*, Yubai Yuan*[2]*, Haoda Fu*[3] *and Annie Qu*[2]. [1]University of Illinois, Urbana-Champaign [2]University of California, Irvine [3]Eli Lilly and Company

18:25(EDT) Transportation of Area Under the ROC curve to a Target Population
♦*Bing Li*[1]*, Issa Dahabreh*[2]*, Constantine Gatsonis*[1] *and Jon Steingrimsson*[1]. [1]Brown University [2]Harvard University

18:50(EDT) Scalable community detection in massive networks via predictive inference
♦*Subhankar Bhadra*[1]*, Marianna Pensky*[2] *and Srijan Sengupta*[1]. [1]NC State University [2]University of Central Florida

19:15(EDT) The Promises of Parallel Outcomes
♦*Ying Zhou, Dehan Kong and Linbo Wang*. University of Toronto

19:40(EDT) A Wavelet-Based Independence Test for Functional Data with an Application to MEG Functional Connectivity
♦*Rui Miao*[1]*, Xiaoke Zhang*[1] *and Raymond Wong*[2]. [1]The George Washington University [2]Texas A&M University

20:05(EDT) Linear Models with Distributed Data and Communication-Efficient Estimator: Limiting Distribution and Its Approximation
Ziyan Yin. Temple University

NA Floor Discussion.

Session 2: Recent Advances in Deep Learning

Organizer: Sijian Wang.
Chair: Sijian Wang.

18:00(EDT) Advances of Momentum in Optimization Algorithm and Neural Architecture Design
Bao Wang. University of Utah

18:25(EDT) Diff-ResNet: Diffusion Augmented Neural Network
Tangjun Wang, Chenglong Bao and ♦*Zuoqiang Shi*. Tsinghua University

18:50(EDT) SODEN: A Scalable Continuous-Time Survival Model through Ordinary Differential Equation Networks
♦*Weijing Tang, Jiaqi Ma, Qiaozhu Mei and Ji Zhu*. University of Michigan

19:15(EDT) A graph deep learning model based on neural ordinary differential equations
♦*Yuanhao Liu, Changpeng Lu and Sijian Wang*. Rutgers University

19:40(EDT) Floor Discussion.

Session 3: Missing Data Method Development and Applications

Organizer: Changchun Xie.
Chair: Changchun Xie.

18:00(EDT) Semiparametrically efficient approach for regression analysis with missing response and a hypothesis testing procedure for the missing data mechanism
♦*Qinglong Tian*[1]*, Jiwei Zhao*[1] *and Donglin Zeng*[2]. [1]University of Wisconsin - Madison [2]The University of North Carolina at Chapel Hill

18:25(EDT) A practical solution for missing data in clinical outcome-a simulation study of multiple imputations in a real-world patient registry
♦*Kim Hung Lo*[1]*, Yiting Wang*[2]*, Chunyan Liu*[3] *and Nanhua Zhang*[3]. [1]Clinical Biostatistics, Janssen R&D [2]Global Epidemiology, Janssen R&D [3]Division of Biostatistics & Epidemiology, Cincinnati Children's Hospital

18:50(EDT) Model-based multiple imputation method for multilevel regression models with left-censored data and derived predictors
♦*Peixin Xu*[1]*, Aimin Chen*[2]*, Nanhua Zhang*[3] *and Changchun Xie*[1]. [1]University of Cincinnati College of Medicine [2]University of Pennsylvania Perelman School of Medicine [3]Cincinnati Children's Hospital Medical Center & University of Cincinnati College of Medicine

19:15(EDT) Floor Discussion.

Session 4: Recent advances in methodologies for omics data analysis

Organizer: Agus Salim.
Chair: Agus Salim.

18:00(EDT) NanoSplicer: Accurate identification of splice junctions using Oxford Nanopore sequencing
Yupei You[1]*, Michael Clark*[2] *and* ♦*Heejung Shim*[1]. [1]School of Mathematics and Statistics and Melbourne Integrative Genomics, The University of Melbourne, Parkville, Australia [2]Centre for Stem Cell Systems, Department of Anatomy and Neuroscience, The University of Melbourne, Parkville, Australia

18:25(EDT) Cross-Platform Omics Prediction procedure: a statistical framework for implementing precision medicine
Jean Yang. The University of Sydney,

18:50(EDT) RUV-III-NB: Removing Unwanted Variation from High-throughput Single-Cell RNA-seq Data
♦*Agus Salim*[1]*, Ramyar Molania*[2]*, Jianan Wang*[2] *and Terence Speed*[2]. [1]University of Melbourne [2]Walter Eliza Hall Institute of Medical Research

19:15(EDT) Multiple Testing of One-sided Hypotheses under General Dependence

Seonghun Cho[1], ♦*Youngrae Kim*[1], *Hyungwon Choi*[2], *Johan Lim*[1], *DoHwan Park*[3] *and Woncheol Jang*[1]. [1]Department of Statistics, Seoul National University [2]School of Medicine, National University of Singapore [3]Department of Mathematics and Statistics, University of Maryland

19:40(EDT) Floor Discussion.

Session 5: Statistical methods for microbiome data analysis

Organizer: Xiaowei Zhan.
Chair: Xiaowei Zhan.

18:00(EDT) mbImpute: an accurate and robust imputation method for microbiome data

♦*Ruochen Jiang*[1], *Wei Vivian Li*[2] *and Jessica Jingyi Li*[1]. [1]University of California, Los Angeles [2]Rutgers school of Public Health

18:25(EDT) Linear Models for Differential Abundance Analysis of Microbiome Compositional Data

♦*Jun Chen*[1] *and Xianyang Zhang*[2]. [1]Mayo Clinic [2]Texas A&M University

18:50(EDT) Variable selection analysis in small-sample microbiome compositional data

Arun Srinivasan, Lingzhou Xue and ♦*Xiang Zhan*. Penn State University

19:15(EDT) Microbial Trend Analysis in Longitudinal Microbiome Study

Chan Wang[1], *Jiyuan Hu*[1], *Martin Blaser*[2] *and* ♦*Huilin Li*[1]. [1]New York University [2]Rutgers University

19:40(EDT) Floor Discussion.

Session 6: Modern Statistical Methods for Analyzing Time Series Data

Organizer: Sumanta Basu.
Chair: Sumanta Basu.

18:00(EDT) Biclustering Approaches for High-Frequency Time Series

Haitao Liu[1], *Jian Zou*[1] *and* ♦*Nalini Ravishanker*[2]. [1]Worcester Polytechnic Institute [2]University of Connecticut

18:25(EDT) Bayesian Estimation of Time-varying models

Sayar Karmakar. University of Florida

18:50(EDT) Spectral methods for small sample time series: A complete periodogram approach

Sourav Das[1], ♦*Suhasini Subba Rao*[2] *and Junho Yang*[3]. [1]James Cook University [2]Texas A&M University [3]Texa A&M University

19:15(EDT) Large Spectral Density Matrix Estimation by Adaptive Thresholding

Yiming Sun[1], *Yige Li*[2], *Amy Kuceyeski*[1] *and* ♦*Sumanta Basu*[1]. [1]Cornell University [2]Harvard University

19:40(EDT) Floor Discussion.

Session 7: Modern machine learning: method and theory

Organizer: Yang Ning.
Chair: Yang Ning.

18:00(EDT) On Function Approximation in Reinforcement Learning: Optimism in the Face of Large State Spaces

♦*Zhuoran Yang*[1], *Chi Jin*[1], *Zhaoran Wang*[2], *Mengdi Wang*[1] *and Michael Jordan*[3]. [1]Princeton [2]Northwestern [3]UC Berkeley

18:25(EDT) The cost of privacy in generalized linear models: algorithms and optimal rate of convergence

Tony Cai[1], *Yichen Wang*[1] *and* ♦*Linjun Zhang*[2]. [1]University of Pennsylvania [2]Rutgers University

18:50(EDT) On the Statistical Properties of Adversarial Robust Estimators

♦*Yue Xing, Qifan Song and Guang Cheng*. Purdue University

19:15(EDT) Floor Discussion.

Session 8: Semiparametric and nonparametric methods in causal inference with multi- or high-dimensional confounders

Organizer: Ming-Yueh Huang.
Chair: Ming-Yueh Huang.

18:00(EDT) Modeling the natural history of human diseases

♦*Yen-Tsung Huang and Ju-Sheng Hong*. Academia Sinica

18:25(EDT) Inference for algorithm-agnostic variable importance

Brian Williamson[1], *Peter Gilbert*[2], *Noah Simon*[3] *and* ♦*Marco Carone*[3]. [1]Kaiser Permanente Washington Health Research Institute [2]Fred Hutchinson Cancer Research Center [3]University of Washington

18:50(EDT) Optimal tests of the composite null hypothesis arising in mediation analysis

♦*Caleb Miles*[1] *and Antoine Chambaz*[2]. [1]Columbia University [2]Université de Paris

19:15(EDT) Statistical methods for improving randomized clinical trial analysis with integrated information from real-world evidencestudies

Shu Yang. North Carolina State University

19:40(EDT) Floor Discussion.

Session 9: Statistics in Biosciences: Recent methodological developments and applications

Organizer: Hongzhe Li.
Chair: Hongzhe Li.

18:00(EDT) Generating Survival Times Using Cox Proportional Hazards Models with Cyclic and Piecewise Time-Varying Covariates

♦*Yunda Huang*[1], *Yuanyuan Zhang*[1], *Zong Zhang*[2] *and Peter Gilbert*[1]. [1]Fred Hutchinson Cancer Center [2]Carnegie Mellon University

18:25(EDT) Assessing Treatment Benefit in Immuno-oncology

Marc Buyse. IDDI

18:50(EDT) Statistical and Computational Approaches for the Identification of Novel Viruses and Virus-host Interactions

Fengzhu Sun. University of Southern California

19:15(EDT) A Hybrid Approach for the Strati¡U+FB01¿ed Mark-Speci¡U+FB01¿c Proportional Hazards Model with Missing Covariates and Missing Marks, with Application to Vaccine Ef¡U+FB01¿cacy Trials
♦*Yanqing Sun*[1]*, Li Qi*[2]*, Fei Heng*[3] *and Peter Gilbert*[4]. [1]The University of North Carolina at Charlotte [2]Sanofi [3]University of North Florida [4]Fred Hutchinson Cancer Research Center and University of Washington)

19:40(EDT) Floor Discussion.

Session 10: New advances in nonparametric and functional data analysis
Organizer: Tiejun Tong.
Chair: Pang Du.

18:00(EDT) Low Rank Approximation for Smoothing Spline via Eigensystem Truncation
♦*Yuedong Wang*[1] *and Danqing Xu*[2]. [1]University of California - Santa Barbara [2]AbbVie

18:25(EDT) A sparse follow-up procedure for functional contrast tests
Quyen Do and ♦*Pang Du*. Virginia Tech

18:50(EDT) Bayesian jackknife empirical likelihood
Yichen Cheng and ♦*Yichuan Zhao*. Georgia State University

19:15(EDT) An application of semiparametric method in Nonparametric Regression
Zhijian Li. UCSB

19:40(EDT) Floor Discussion.

Session 11: Recent advances in statistical methods for complex biomedical data
Organizer: Yuehua Cui.
Chair: Yuehua Cui.

18:00(EDT) Decomposing cell compositional effect in bulk tissue gene expression analysis
♦*Shaoyu Li*[1]*, Xue Wang*[2]*, Duan Chen*[1] *and Jinjing Zhang*[1]. [1]University of North Carolina [2]Mayo Clinics

18:25(EDT) Robust Estimation of Heterogeneous Treatment Effects using Electronic Health Record Data
Ruohong Li[1]*,* ♦*Honglang Wang*[2] *and Wanzhu Tu*[1]. [1]Indiana University [2]Indiana University-Purdue University Indianapolis

18:50(EDT) A Fast and Efficient Likelihood Approach for Genome-wide Mediation Analysis
Janaka Liyanage[1]*, Jeremie Estepp*[1]*, Kumar Srivastava*[1]*, Yun Li*[2]*, Motomi Mori*[1] *and* ♦*Guolian Kang*[1]. [1]St. Jude Children's Research Hospital [2]The University of North Carolina at Chapel Hill

19:15(EDT) Multi-marker survival tests for interval censored data
♦*Chenxi Li*[1]*, Di Wu*[1] *and Qing Lu*[2]. [1]Michigan State University [2]University of Florida

19:40(EDT) Floor Discussion.

Session 12: Recent Advances in Causal Inference With Applications to Public Health and Policy
Organizer: Xu Shi.
Chair: Kendrick Li.

18:00(EDT) Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment
Kosuke Imai[1]*,* ♦*Zhichao Jiang*[2]*, James Greiner*[3]*, Ryan Halen*[3] *and Sooahn Shin*[1]. [1]Harvard University [2]University of Massachusetts Amherst [3]Harvard Law School

18:25(EDT) Propensity score weighting analysis of survival outcomes using pseudo-observations
Shuxi Zeng[1]*, Fan Li*[1]*, Liangyuan Hu*[2] *and* ♦*Fan Li*[3]. [1]Duke University [2]Rutgers University [3]Yale University

18:50(EDT) A negative correlation strategy for bracketing in difference-in-differences
♦*Ting Ye*[1]*, Luke Keele*[1]*, Raiden Hasegawa*[2] *and Dylan Small*[1]. [1]University of Pennsylvania [2]Google

19:15(EDT) Floor Discussion.

Session 13: New Statistical Methods for Competing Risks
Organizer: Xuewen Lu.
Chair: Xuewen Lu.

18:00(EDT) Doubly Robust Estimation of the Hazard Difference for Competing Risks Data
♦*Denise Rava and Ronghui (Lily) Xu*. UC San Diego

18:25(EDT) Variable selection in semiparametric transformation regression with interval-censored competing risks data
♦*Fatemeh Mahmoudi and Xuewen Lu*. University of Calgary

18:50(EDT) Semiparametric Marginal Regression for Clustered Competing Risks Data with Missing Cause of Failure
Wenxian Zhou[1]*,* ♦*Giorgos Bakoyannis*[1]*, Ying Zhang*[2] *and Constantin Yiannoutsos*[1]. [1]Indiana University [2]University of Nebraska Medical Center

19:15(EDT) Joint Correlation Learning for Semi-competing Risks Outcomes with Ultrahigh Dimensional Covariates
Mengjiao Peng[1] *and* ♦*Liming Xiang*[2]. [1]East China Normal University, Shanghai [2]Nanyang Technological University, Singapore

19:40(EDT) Floor Discussion.

Session 14: Recent Advancements in Large-Scale and High-dimensional Inference
Organizer: Xianyang Zhang.
Chair: Xianyang Zhang.

18:00(EDT) Individual data protected meta-analysis of large scale healthcare data from multiple sites
Tianxi Cai[1]*,* ♦*Molei Liu*[1] *and Yin Xia*[2]. [1]Harvard Chan School of Public Health [2]Fudan University

18:25(EDT) Tuning-free large-scale multivariate regression via shrinkage estimation
Yihe Wang[1] *and* ♦*Dave Zhao*[2]. [1]Facebook, Inc. [2]University of Illinois Urbana-Champaign

18:50(EDT) LinDA: Linear Models for Differential Abundance Analysis of Microbiome Compositional Data
♦*Huijuan Zhou*[1]*, Xianyang Zhang*[2]*, Kejun He*[3] *and Jun Chen*[4]. [1]Texas A&M University; Renmin University of China [2]Texas A&M University [3]Renmin University of China [4]Mayo Clinic

19:15(EDT) A flexible model-free prediction-based framework for feature ranking
♦*Jingyi Jessica Li*[1], *Yiling Chen*[1] *and Xin Tong*[2]. [1]University of California Los Angeles [2]University of Southern California

19:40(EDT) Floor Discussion.

Sep. 13 10:20-noon (EDT)

Session 15: Statistical Learning and Variable Selection

Organizer: Wei Zhong.
Chair: Wei Zhong.

10:20(EDT) A Tweedie Compound Poisson Model in RKHS
Yi Yang. McGill University

10:45(EDT) Simultaneous Variable Selection and Covariance Estimation in High Dimensional Linear Mixed Model
Xin Liu. Shanghai University of Finance and Economics

11:10(EDT) Nonconvex clustering with random projection
Xiaodong Yan. Shandong University

11:35(EDT) Bootstrap Inference for the Finite Population Mean under Complex Sampling Designs
♦*Zhonglei Wang*[1], *Liuhua Peng*[2] *and Jae-Kwang Kim*[3]. [1]Xiamen University [2]The University of Melbourne [3]Iowa State University

12:00(EDT) Floor Discussion.

Session 16: Recent advances in treatment evaluation and risk prediction with survival data

Organizer: Chen Hu.
Chair: Chen Hu.

10:20(EDT) Dynamic risk prediction of adverse outcomes of ICU patients with sepsis using machine learning methods
Chung-Chou H. Chang. University of Pittsburgh

10:45(EDT) On restricted mean time in favor of treatment
Lu Mao. N/A

11:10(EDT) The benefit-risk assessment of new treatments using generalized pairwise comparisons, a population- and individual-level perspective
julien peron. hospices de civils de lyon

11:35(EDT) Floor Discussion.

Session 17: Statistics in Microbiome Research

Organizer: Anru Zhang.
Chair: Pixu Shi.

10:20(EDT) Microbiome multi-omics integration using compositional reduced rank regression for relative abundance data
Chenxiao Hu and ♦*Duo Jiang*. Oregon State University

10:45(EDT) Measurement error in compositional data and the replicability of microbiome studies
Amy Willis. University of Washington

11:10(EDT) Dimension Reduction of Longitudinal Microbiome Data via Tensor Functional SVD
♦*Pixu Shi, Rungang Han and Anru Zhang*. Duke University

11:35(EDT) Joint modeling of zero-inflated longitudinal proportions and time-to-event data with application to a gut microbiome study
♦*Jiyuan Hu*[1], *Chan Wang*[1], *Martin Blaser*[2] *and Huilin Li*[1]. [1]NYU Grossman School of Medicine [2]Rutgers University

12:00(EDT) Floor Discussion.

Session 18: Statistical Text Mining

Organizer: Tracy Ke.
Chair: Tracy Ke.

10:20(EDT) How Much Can Machines Learn Finance From Chinese Text Data?
♦*Jianqing Fan, Lirong Xue and Yang Zhou*. Princeton University

10:45(EDT) Likelihood estimation of sparse topic distributions in topic models and its applications to Wasserstein document distance calculations
♦*Florentina Bunea, Mike Bing, Marten Wegkamp and Seth Strimas-Mackey*. Cornell University

11:10(EDT) Topic Analysis of Statistical Abstracts
Pengsheng Ji[1], *Jiashun Jin*[2], ♦*Tracy Ke*[3] *and Wanshan Li*[2]. [1]University of Georgia [2]Carnegie Mellon University [3]Harvard University

11:35(EDT) Floor Discussion.

Session 19: Recent Developments on Network Data Analysis

Organizer: Tracy Ke.
Chair: Jiaming Xu.

10:20(EDT) Optimal adaptivity of Signed-Polygon for network testing
Jiashun Jin. Carnegie Mellon University

10:45(EDT) Fast Network Community Detection with Profile-Pseudo Likelihood Methods
Ji Zhu. University of Michigan

11:10(EDT) Testing correlation of unlabeled random graphs
Yihong Wu[1], ♦*Jiaming Xu*[2] *and Sophie H. Yu*[2]. [1]Yale University [2]Duke University

11:35(EDT) Recommendation system with social network data by tensor factorization
Yiyuan Liu, Ya Wang and ♦*Bingyi Jing*. HKUST

12:00(EDT) Floor Discussion.

Session 20: Causal inference methods: propensity score, matching, imputation and more

Organizer: Yun Li.
Chair: Di Shu.

10:20(EDT) Propensity score analysis methods with balancing constraints: a Monte Carlo study
Yan Li and ♦*Liang Li*. MD Anderson Cancer Center

10:45(EDT) Estimating Causal Effect of Restricted Mean Survival Time Difference based on Matched Design in Observational Survival Data
♦*Zihan Lin, Andy Ni and Bo Lu*. OSU

11:10(EDT) Using multiple imputation to classify potential outcomes subgroups
Yun Li. University of Pennsylvania

11:35(EDT) Regression-based causal inference with factorial experiments: estimands, model specifications, and design-based properties
Anqi Zhao[1] *and* ♦*Peng Ding*[2]. [1]National University of Singapore [2]University of California Berkeley

12:00(EDT) Floor Discussion.

Session 21: Trial design and efficacy estimation of COVID-19 vaccine
Organizer: Yun Li.
Chair: Ying Huang.

10:20(EDT) Trial design and efficacy estimation of COVID-19 vaccine
Peter Gilbert. Fred Hutchinson Cancer Research Center

10:45(EDT) Evaluating Vaccine Efficacy Against SARS-CoV-2 Infection
Danyu Lin. University of North Carolina

11:10(EDT) Identifying COVID-19 vaccine correlates of protection from randomized trials
♦*David Benkeser*[1], *Iván Díaz*[2] *and Jialu Ran*[3]. [1]Emory University [2]Weill Cornell [3]jialu.ran@emory.edu

11:35(EDT) Vaccine Development During Pandemic: Innovative Design Transforming Development Paradigm
Satrajit Roychoudhury. Pfizer Inc

12:00(EDT) Floor Discussion.

Session 22: Manifold Learning and Geometric Methods in Statistics
Organizer: Wanli Qiao.
Chair: Wanli Qiao.

10:20(EDT) Principal Sub-manifolds and Classification on Manifolds
Zhigang Yao. National University of Singapore

10:45(EDT) Intrinsic and extrinsic deep learning on manifolds
♦*Lizhen Lin*[1], *Yihao Fang*[1], *Ilsang Ohn*[1] *and Bayan Saparbayeva*[2]. [1]The University of Notre Dame [2]The University of Rochester

11:10(EDT) Gaussian process subspace regression: How to do PCA without a data sample?
♦*Ruda Zhang, Simon Mak and David Dunson*. Duke University

11:35(EDT) Amplitude mean of trajectories on S2
Zhengwu Zhang. UNC Chapel Hill

12:00(EDT) Floor Discussion.

Session 23: Modern streaming Data Analysis: detection and identification
Organizer: Ruizhi Zhang.
Chair: Ruizhi Zhang.

10:20(EDT) Industrial Analytics Research Facing Digital Transformation
Fugee Tsung. HKUST

10:45(EDT) An EWMA chart for High dimensional Process with Multi-class Out-of-Control Information via Random Forest Learning
Dongdong Xiang. East China Normal University

11:10(EDT) Adaptive Process Monitoring Using Covariate Information
Kai Yang[1] *and* ♦*Peihua Qiu*[2]. [1]Graduate Assistant [2]Professor and Founding Chair

11:35(EDT) Item Pool Quality Control in Educational Testing via Compound Sequential Change Detection
Yunxiao Chen[1], *Yi-Hsuan Lee*[2] *and* ♦*Xiaoou Li*[3]. [1]London School of Economics and Political Science [2]Educational Testing Service [3]University of Minnesota

12:00(EDT) Floor Discussion.

Session 24: Robust methods for feature selection in high-dimensional problems
Organizer: David Kepplinger.
Chair: David Kepplinger.

10:20(EDT) The Bayesian Regularized Quantile Varying Coefficient Model
♦*Cen Wu*[1], *Fei Zhou*[1] *and Jie Ren*[2]. [1]Department of Statistics, Kansas State University [2]Department of Biostatistics and Health Data Sciences, Indiana University School of Medicine

10:45(EDT) Recent developments in robust estimation and variable selection procedures
Irène Gijbels Gijbels. KU Leuven, Belgium

11:10(EDT) Robust Regression With Covariate Filtering: Heavy Tails and Adversarial Contamination
Ankit Pensia[1], *Varun Jog*[2] *and* ♦*Po-Ling Loh*[2]. [1]UW-Madison [2]University of Cambridge

11:35(EDT) Simultaneous Feature Selection and Outlier Detection with Optimality Guarantees
♦*Luca Insolia*[1], *Ana Kenney*[2], *Francesca Chiaromonte*[2] *and Giovanni Felici*[3]. [1]Sant'Anna School of Advanced Studies [2]Pennsylvania State University [3]National Research Council of Italy

12:00(EDT) Floor Discussion.

Session 25: Nonparametric Learning: New Directions and Innovations
Organizer: Lily Wang.
Chair: Guannan Wang.

10:20(EDT) Multiple domain and multiple kernel outcome-weighted learning for estimating individualized treatment regimes
Shanghong Xie[1], *Thaddeus Tarpey*[2], *Eva Petkova*[2] *and* ♦*Todd Ogden*[1]. [1]Columbia University [2]NYU

10:45(EDT) Cross-sectional correlation tests with functional data and its application to inter- subject correlation analysis
Hongnan Wang and ♦*Ping-Shou Zhong*. University of Illinois at Chicago

11:10(EDT) Sparse Modeling of Functional Linear Regression via Fused Lasso with Application to Genotype-by-environment Interaction Studies

♦*Shan Yu*[1]*, Aaron Kusmec*[2]*, Li Wang*[3] *and Dan Nettleton*[3]. [1]University of Virginia [2]George Mason University [3]Iowa State University

11:35(EDT) Partial Separability and Functional Graphical Models

Javier Zapata[1]*, Sang-Yun Oh*[1] *and* ♦*Alexander Petersen*[2]. [1]University of California Santa Barbara [2]Brigham Young University

12:00(EDT) Floor Discussion.

Session 26: Challenges in the analysis of complex structured data

Organizer: Wenqing He.
Chair: Wenqing He.

10:20(EDT) Cox regression with dynamic markers measured at covariate-dependent visit times

♦*Richard Cook and Bingfeng Xie*. University of Waterloo

10:45(EDT) Semiparametric Regression Model for Ordinal Response

Ao Yuan. Georgetown University

11:10(EDT) Risk Projection for Time-to-Event Leveraging Summary Statistics With Source Individual-level Data

Jiayin Zheng, Yingye Zheng and ♦*Li Hsu*. Fred Hutchinson Cancer Research Center

11:35(EDT) A multiple imputation method for nonlinear mixed effects models with missing data

Lang Wu. University of British Columbia

12:00(EDT) Floor Discussion.

Session 27: Recent development on the analysis of single-cell RNA-seq data

Organizer: Yingying Wei.
Chair: Yingying Wei.

10:20(EDT) Flexible experimental designs for valid single-cell RNA-sequencing experiments allowing batch effects correction

♦*Fangda Song*[1] *and Yingying Wei*[2]. [1]The Chinese University of Hong Kong, Shenzhen [2]The Chinese University of Hong Kong

10:45(EDT) Model-Based Trajectory Inference for Single-Cell RNA Sequencing Using Deep Learning with a Mixture Prior

Jinhong Du, Ming Gao and ♦*Jingshu Wang*. University of Chicago

11:10(EDT) Estimating Heterogeneous Gene Regulatory Networks from Zero-Inflated Single-Cell Expression Data

♦*Xiangyu Luo and Qiuyu Wu*. Renmin University of China

11:35(EDT) Non-Parametric Modeling Enables Scalable and Robust Detection of Spatial Expression Patterns for Large Spatial Transcriptomic Studies

Xiang Zhou. University of Michigan

12:00(EDT) Floor Discussion.

Session 28: AI Boosting of pharmaceutical drug development and the emerging world of digital therapeutics

Organizer: Margaret Gamalo.
Chair: Margaret Gamalo.

10:20(EDT) Subgroup analysis based on K-means for multivariate longitudinal data

♦*Gen Zhu*[1]*, Margaret Gamalo*[2] *and Chuanbo Zang*[2]. [1]UTHealth [2]Pfizer

10:45(EDT) Beyond step counting-Measuring Nighttime Scratch and Sleep with Wearable Devices

Nikhil Mahadevan, Yiorgos Christakis, ♦*Junrui Di, Jonathan Bruno, Yao Zhang, Carrie Northcott and Shyamal Pate*. Pfizer

11:10(EDT) Statistical Considerations for the Clinical Validations of AI/ML-based Digital Health Products

Feiming Chen. Food and Drug Administration

11:35(EDT) Floor Discussion.

Sep. 13 2-3:40pm(EDT)

Session 29: Real World Evidence Innovation using Causal Methodology and Application

Organizer: Yi Huang.
Chair: Martin Klein.

14:00(EDT) Propensity Score Methods in for Multiple Treatment Comparisons for Evaluating Drug Safety and Effectiveness

Rongmei Zhang. Center for Drug Evaluation and Research

14:25(EDT) Comparison Study of Causal Methods for Average Treatment Effect Estimation Allowing Covariate Measurement Error

♦*Yi Huang*[1] *and Feng Zhou*[2]. [1]UMBC [2]FDA

14:50(EDT) Real World Data for Clinical Development: Clinical Eligibility Criteria, Hybrid Control Arms, and Challenges

Thomas Jemielita. Merck & Co., Inc.

15:15(EDT) Optimal Treatment Regimes: An Empirical Comparison of Methods and Applications

Zhen Li[1]*,* ♦*Jie Chen*[2]*, Eric Labor*[3]*, Fang Liu*[4] *and Richard Baumgartner*[4]. [1]Amazon [2]Overland Pharma [3]Duke University [4]Merck

15:40(EDT) Floor Discussion.

Session 30: Recent development in Pediatric Clinical Trial Design

Organizer: Ran Duan.
Chair: James Travis.

14:00(EDT) Causal Inference in Rare Diseases

Rima Izem. Novartis

14:25(EDT) The Potential of Master Protocols in Pediatric Clinical Trials

Kristine Broglio. AstraZeneca

14:50(EDT) Integrating adult data into design of pediatric dose-finding studies

♦*Yimei Li*[1] *and Ying Yuan*[2]. [1]University of Pennsylvania [2]University of Texas MD Anderson Cancer Center

15:15(EDT) Floor Discussion.

Session 31: Recent development in machine learning and causal inference
Organizer: Yang Ning.
Chair: Yang Ning.

14:00(EDT) Conformal Inference of Counterfactuals and Individual Treatment Effects
♦*Lihua Lei and Emmanuel Candès*. Stanford University

14:25(EDT) Optimal and Safe Estimation for High-Dimensional Semi-Supervised Learning
Jiwei Zhao. University of Wisconsin Madison

14:50(EDT) Doubly Robust Semiparametric Inference Using Regularized Calibrated Estimation with High-dimensional Data
Zhiqiang Tan. Rutgers University

15:15(EDT) Profile Matching for the Generalization and Personalization of Causal Inferences
Eric Cohn and ♦*Jose Zubizarreta*. Harvard University

15:40(EDT) Floor Discussion.

Session 32: Prediction and inference for neural-network-based methods and their applications to genetics
Organizer: Pei Geng.
Chair: Pei Geng.

14:00(EDT) A New Kernel Neural Network Method for Genetic Data Analysis
♦*Qing Lu*[1], *Xiaoxi Shen*[1] *and Xiaoran Tong*[2]. [1]University of Florida [2]NIH

14:25(EDT) Expectile Neural Networks for Genetic DataAnalysis of Complex Diseases
♦*Jinghang Lin*[1], *Xiaoran Tong*[2], *Chenxi Li*[1] *and Qing Lu*[3]. [1]Michigan State University [2]National Institutes of Health [3]University of Florida

14:50(EDT) A Goodness-of-Fit Test Based on Neural Networks
Xiaoxi Shen and ♦*Chang Jiang*. University of Florida

15:15(EDT) Multimodal Functional Deep Learning for Multi-omics Data
♦*Yuan Zhou*[1], *Shan Zhang*[2], *Pei Geng*[3] *and Qing Lu*[1]. [1]University of Florida [2]Michigan State University [3]Illinois State University

15:40(EDT) Floor Discussion.

Session 33: The fad and fade in statistics for biopharmaceutical research: editor?s pick from top biopharmaceutical statistics journals
Organizer: Junjing Lin.
Chair: Jianchang Lin.

14:00(EDT) BIOSTATISTICS RESEARCH TRENDS IN BIOMETRICS, THE FLAGSHIP JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY, IN PRE-, PERI-, AND POST-PANDEMIC TIMES
Geert Molenberghs. UHasselt & KU Leuven

14:25(EDT) Not-so-popular stories - statistical methods that are not in the mainstream of biopharmaceutical statistics literature
Margaret Gamalo. Pfizer

14:50(EDT) Statistics in Medicine: what's new and what's not new?
Robert Platt. McGill University

15:15(EDT) Out of its infancy: Statistics in Biopharmaceutical Research
♦*Frank Bretz*[1] *and Toshimitsu Hamasaki*[2]. [1]Novartis Pharma AG [2]George Washington University

15:40(EDT) Floor Discussion.

Session 34: Recent Development in Multi-Block Data Integration
Organizer: Lu Xia.
Chair: Lu Xia.

14:00(EDT) Simultaneous Clustering and Estimation of Networks via Sparse Tensor Decomposition
♦*Gen Li*[1] *and Miaoyan Wang*[2]. [1]University of Michigan, Ann Arbor [2]University of Wisconsin, Madison

14:25(EDT) Semi-Supervised Statistical Inference for High-Dimensional Linear Regression with Blockwise Missing Data
♦*Fei Xue*[1], *Rong Ma*[2] *and Hongzhe Li*[2]. [1]Purdue University [2]University of Pennsylvania

14:50(EDT) Joint Matrix Decomposition Regression for Multi-omics Data Analysis
♦*Yue Wang*[1], *Jing Ma*[2], *Timothy Randolph*[2] *and Ali Shojaie*[3]. [1]Arizona State University [2]Fred Hutch Cancer Research Center [3]University of Washington

15:15(EDT) A Decomposition-based Canonical Correlation Analysis for High-Dimensional Datasets
♦*Hai Shu*[1], *Xiao Wang*[2] *and Hongtu Zhu*[3]. [1]New York University [2]Purdue University [3]The University of North Carolina at Chapel Hill

15:40(EDT) Floor Discussion.

Session 35: Modern streaming Data Analysis: process monitoring
Organizer: Ruizhi Zhang.
Chair: Yajun Mei.

14:00(EDT) Detection and identification
Olympia Hadjiliadis. Hunter College

14:25(EDT) Hot-spot detection and identification for Poisson data
Yujie Zhao, Xiaoming Huo and ♦*Yajun Mei*. Georgia Institute of Technology

14:50(EDT) Adaptive Partially-Observed Sequential Change Point Detection for Covid-19 hotspots detection
Jiuyun Hu[1], ♦*Hao Yan*[1], *Yanjun Mei*[2] *and Sarah Holte*[3]. [1]Arizona State University [2]Georgia Institution of Technology [3]University of Washington

15:15(EDT) Efficient Global Monitoring Statistics for High-Dimensional Data
Jun Li. University of California, Riverside

15:40(EDT) Floor Discussion.

Session 36: Advances in Spatio-Temporal Statistics
Organizer: Ying Sun.
Chair: Ying Sun.

14:00(EDT) Second-order semi-parametric inference for multivariate log Gaussian Cox processes
Kristian Hessellund[1], ♦*Ganggang Xu*[2], *Yongtao Guan*[2] *and Rasmus Waagepetersen*[1]. [1]Aalborg University [2]University of Miami

14:25(EDT) Skew-Elliptical Cluster Processes
♦*Ngoc Anh Dao*[1] *and Marc Genton*[2]. [1]Texas A&M University [2]King Abdullah University of Science and Technology

14:50(EDT) Functional singular spectrum analysis
Hossein Haghbin[1], *S. Morteza Najibi*[2], *Rahim Mahmoudvand*[3], *Jordan Trinka*[4] *and* ♦*Mehdi Maadooliat*[5]. [1]Persian Gulf University [2]Lund University [3]Bu-Ali Sina University [4]Pacific Northwest National Laboratory [5]Marquette University

15:15(EDT) Bayesian Co-kriging Model for Remote Sensing Measurements with Different Quality Flags: Uncertainty Quantification in NASA's AIRS Mission
Bledar Konomi. University of Cincinnati

15:40(EDT) Floor Discussion.

Session 37: Advances in Spatial and Spatio-temporal Modeling and its Applications

Organizer: Seiyon Ben Lee.
Chair: Seiyon Ben Lee.

14:00(EDT) A combined physical-statistical approach for estimating storm surge risk
♦*Whitney Huang*[1] *and Emily Tidwell*[2]. [1]Clemson University [2]Dynetics

14:25(EDT) Scalable Forward Sampler Backward Smoother based on the Vecchia approximation
♦*Marcin Jurek*[1], *Matthias Katzfuss*[2] *and Pulong Ma*[3]. [1]University of Texas at Austin [2]Texas A&M University [3]Duke University

14:50(EDT) Conjugate spatio-temporal Bayesian multinomial Polya-gamma regression for the reconstruction of climate using pollen
John Tipton. 1700 N Old Wire Rd

15:15(EDT) Hierarchical Integrated Spatial Process Modeling of Monotone West Antarctic Snow Density Curves
♦*Philip White*[1], *Durban Keeler*[2] *and Summer Rupper*[2]. [1]Brigham Young University [2]University of Utah

15:40(EDT) Floor Discussion.

Session 38: Recent Challenges and Advances for Complex Survival Data

Organizer: Hua Shen.
Chair: Hua Shen.

14:00(EDT) Censored quantile regression with auxiliary information
Zhaozhi Fan. Memorial University

14:25(EDT) A comparative study of R packages for semiparametric shared frailty models
Tingxuan Wu[1], *Longhai Li*[1] *and* ♦*Cindy Feng*[2]. [1]University of Saskatchewan [2]cindy.feng@dal.ca

14:50(EDT) Accelerated Failure Time Models for Joint Analysis of Longitudinal and Time-to-Event Data
Shahedul Khan. University of Saskatchewan

15:15(EDT) Cox Proportional-Hazards Regression with Surrogates for Categorical Covariate and Left-Truncated Data
Hua Shen. University of Calgary

15:40(EDT) Floor Discussion.

Session 39: COVID-19 Modeling, Projection and Inference

Organizer: Lily Wang.
Chair: Lily Wang.

14:00(EDT) Dynamic COVID risk assessment accounting for community virus exposure from a spatial-temporal transmission model
Yuanjia Wang. Columbia University

14:25(EDT) Better Strategies for Containing COVID-19 Pandemic–A Study of 25 Countries via a vSIADR Model
Yumou Qiu. Iowa State University

14:50(EDT) Space-time Epidemic Models: The Complexity of Simplicity
Myungjin Kim[1], *Zhiling Gu*[2], *Shan Yu*[3], ♦*Guannan Wang*[4] *and Li Wang*[5]. [1]Brown University [2]Iowa State University [3]University of Virginia [4]College of William & Mary [5]George Mason university

15:15(EDT) Analyzing COVID-19 Data: Some Issues and Challenges
Grace Yi. University of Western Ontario

15:40(EDT) Floor Discussion.

Session 40: Design and Analysis Experiments for Developing Mobile Health Devices

Organizer: Xiaobo Zhong.
Chair: Yutao Liu.

14:00(EDT) Building health application recommender system using partially penalized regression
♦*Eun Jeong Oh*[1], *Min Qian*[2], *Ken Cheung*[2] *and David Mohr*[3]. [1]University of Pennsylvania [2]Columbia University [3]Northwestern University

14:25(EDT) A Zero-Inflated Mixture Model for Multivariate Data Clustering
Bin Cheng, ♦*Xiaoqi Lu and Ying-Kuen Cheung*. Columbia University

14:50(EDT) Apply adaptive randomization to sequential multiple assignment randomized trial to develop a smartphone apps platform for depression management
Ying Kuen Cheung[1], ♦*Xiaobo Zhong*[2] *and Bibhas Chakraborty*[3]. [1]Columbia University [2]Bristol Myers Squibb [3]National University of Singapore

15:15(EDT) Personalized Policy Learning using Longitudinal Mobile Health Data
♦*Xinyu Hu*[1], *Min Qian*[2], *Bin Cheng*[2] *and Ying Kuen Cheung*[2]. [1]Uber [2]Columbia University

15:40(EDT) Floor Discussion.

Session 41: Recent Advances in Statistical Inference for Discrete Structures

Organizer: Anderson Ye Zhang.
Chair: Anderson Ye Zhang.

14:00(EDT) Global and Individualized Community Detection in Inhomogeneous Multilayer Networks
Shuxiao Chen[1], *Sifan Liu*[2] *and* ♦*Zongming Ma*[1]. [1]University of Pennsylvania [2]Stanford University

14:25(EDT) Exact Clustering in Tensor Block Model: Statistical Optimality and Computational Limit
Anru Zhang. Duke University

14:50(EDT) Hierarchical stochastic block model for community detection in multiplex networks
♦*Arash Amini*[1], *Marina Paez*[2] *and Lizhen Lin*[3]. [1]UCLA [2]Federal University of Rio de Janeiro [3]University of Notre Dame

15:15(EDT) Maximum likelihood for high-noise group orbit estimation and single-particle cryo-EM
♦*Zhou Fan*[1], *Roy R. Lederman*[1], *Yi Sun*[2], *Tianhao Wang*[1] *and Sheng Xu*[1]. [1]Yale University [2]University of Chicago

15:40(EDT) Floor Discussion.

Sep. 13 4-5:40pm(EDT)

Session 42: Advances in Detection of Change Points and Signals

Organizer: Dan Cheng.
Chair: Yunpeng Zhao.

16:00(EDT) Multiple Changepoint Detection for Time Series Data
Robert Lund. University of California, Santa Cruz

16:25(EDT) Detection and estimation of local signals
♦*Xiao Fang*[1], *Jian Li*[2] *and David Siegmund*[3]. [1]The Chinese University of Hong Kong [2]Adobe [3]Stanford University

16:50(EDT) Minimax change point testing in the high-dimensional regression setting
Zifeng Zhao[1], *Alessandro Rinaldo*[2] *and* ♦*Daren Wang*[1]. [1]Notre Dame [2]CMU

17:15(EDT) Multiple Testing of Local Extrema for Detection of Change Points
♦*Dan Cheng*[1], *Zhibing He*[1] *and Armin Schwartzman*[2]. [1]Arizona State University [2]University of California San Diego

17:40(EDT) Floor Discussion.

Session 43: Recent development in high-dimensional statistics and machine learning

Organizer: Yang Ning.
Chair: Yang Ning.

16:00(EDT) Statistical Inference Using Conformal Prediction
Jing Lei. Carnegie Mellon University

16:25(EDT) Adaptive Estimation of Multivariate Regression with Hidden Variables
Yang Ning. Cornell

16:50(EDT) Augmented Direct Learning for Conditional Average Treatment Effect Estimation with Double Robustness
Haomiao Meng[1] *and* ♦*Xingye Qiao*[2]. [1]Amazon [2]Binghamton University

17:15(EDT) Efficient Learning of Optimal Individualized Treatment Rules
Weibin Mo and ♦*Yufeng Liu*. University of North Carolina at Chapel Hill

17:40(EDT) Floor Discussion.

Session 44: Historical Controls for Medical Product Development

Organizer: Elande Baro.
Chair: Elande Baro.

16:00(EDT) Historical Borrowing in Pediatric Clinical Trials: An Overview and Regulatory Perspective
James Travis. NA

16:25(EDT) Data-Adaptive Weighting of Real-World and Randomized Controls Using Propensity Scores: Creating a Hybrid Control Arm
♦*Joanna Harton, Rebecca Hubbard and Nandita Mitra*. University of Pennsylvania

16:50(EDT) NA
Anthony J. Hatswell. NA

17:15(EDT) Floor Discussion.

Session 45: New machine learning methods using subgrouping

Organizer: Annie Qu.
Chair: Fei Xue.

16:00(EDT) Dynamic tensor recommender systems
Yanqing Zhang[1], ♦*Xuan Bi*[2], *Niansheng Tang*[1] *and Annie Qu*[3]. [1]Yunnan University [2]University of Minnesota [3]University of California, Irvine

16:25(EDT) Crowdsourcing Utilizing Subgroup Structure of Latent Factor Modeling
♦*Qi Xu*[1], *Yubai Yuan*[1], *Junhui Wang*[2] *and Annie Qu*[1]. [1]UC Irvine [2]City University of Hong Kong

16:50(EDT) A Tensor Factorization Recommender System with Dependency
♦*Jiuchen Zhang*[1], *Yubai Yuan*[2] *and Annie Qu*[3]. [1]PhD student [2]PhD [3]Chancellor's Professor

17:15(EDT) Floor Discussion.

Session 46: Recent advances in the analysis of complex event time data

Organizer: Yifei Sun.
Chair: Yifei Sun.

16:00(EDT) Semiparametric regression analysis of bivariate censored events in a family study of Alzheimer's disease
♦*Fei Gao*[1], *Donglin Zeng*[2] *and Yuanjia Wang*[3]. [1]Fred Hutchinson Cancer Research Center [2]University of North Carolina at Chapel Hill [3]Columbia University

16:25(EDT) Censored Linear Regression in the Presence or Absence of Auxiliary Survival Information
Ying Sheng. University of California at San Francisco

16:50(EDT) Recurrent event trees and ensembles
♦*Yifei Sun*[1], *Sy Han (Steven) Chiou*[2] *and Chiung-Yu Huang*[3]. [1]Columbia University [2]University of Texas at Dallas [3]University of California San Francisco

17:15(EDT) Statistical methods for semi-competing risks modeling with application to SEER-Medicare data
♦*Hong Zhu*[1], *Yu Lan*[2], *JIng Ning*[3] *and Yu Shen*[3]. [1]The University of Texas Southwestern Medical Center [2]Southern Methodist University [3]The University of Texas MD Anderson Cancer Center,

17:40(EDT) Floor Discussion.

Session 47: Emerging Topics in Statistical Genetics and Genomics
Organizer: Li-Xuan Qin.
Chair: Jaya Satagopan.

16:00(EDT) Efficient SNP-based Heritability Estimation using Gaussian Predictive Process in Large-scale Cohort Studies
Souvik Seal[1], *Abhirup Datta*[2] *and* ♦*Saonli Basu*[3]. [1]Colorado School of Public Health [2]Johns Hopkins University [3]University of Minnesota

16:25(EDT) Hurdle Poisson Model-based Clustering for Microbiome Data
Zhili Qiao and ♦*Peng Liu*. Iowa State University

16:50(EDT) Sparsity in Single Cell Hi-C Data-Not All Zeros Are Created Equal
Shili Lin. Ohio State University

17:15(EDT) Model-based analysis of alternative polyadenylation using 3' end reads
♦*Wei Vivian Li*[1], *Dinghai Zheng*[1], *Ruijia Wang*[1] *and Bin Tian*[2]. [1]Rutgers, The State University of New Jersey [2]Wistar Institute

17:40(EDT) Floor Discussion.

Session 48: Statistical considerations for master protocol in I-O and Cell Therapy
Organizer: Rachael Liu.
Chair: Rachael Liu.

16:00(EDT) Statistical considerations for master protocol in I-O and Cell Therapy
Yuan Ji. The University of Chicago

16:25(EDT) Statistical considerations for master protocol in I-O and Cell Therapy
Jianchang Lin[1], *Yuan Ji*[2], ♦*Peter Thall*[3] *and Shentu Yue*[4]. [1]Takeda Pharmaceuticals [2]The University of Chicago [3]The University of Texas MD Anderson Cancer Center [4]Merck

16:50(EDT) Model-based inference with nonconcurrent control in Platform Trials
Anqi Pan[1], *Nicole Li*[2], *Thomas Jemielita*[2] *and* ♦*Yue Shentu*[2]. [1]School of Public Health, University of Georgia [2]Merck & Co

17:15(EDT) Floor Discussion.

Session 49: Recent Developments in Resampling Techniques
Organizer: Rohit Patra.
Chair: Rohit Patra.

16:00(EDT) Asymptotics of cross-validation and the bootstrap.
Morgane Austern. Harvard

16:25(EDT) Bootstrap-Assisted Inference for Generalized Grenander-type Estimators
♦*Matias Cattaneo*[1], *Michael Jansson*[2] *and Kenichi Nagasawa*[3]. [1]Princeton University [2]UC-Berkeley [3]University of Warwick

16:50(EDT) Dependence-Robust Inference Using Resampled Statistics
Michael Leung. UCSC

17:15(EDT) On Gaussian Approximation for M-Estimator
♦*Masaaki Imaizumi*[1] *and Taisuke Otsu*[2]. [1]The University of Tokyo [2]London School of Economics

17:40(EDT) Floor Discussion.

Session 50: Modern techniques in statistical learning
Organizer: Linglong Kong.
Chair: Linglong Kong.

16:00(EDT) Deep Neural Network with a Smooth Monotonic Output Layer for Dynamic Risk Prediction
♦*Zhiyang Zhou*[1], *Yu Deng*[1], *Lei Liu*[2], *Hongmei jiang*[3], *Yifan Peng*[4], *Xiaoyun Yang*[1], *Yun Zhao*[5], *Hongyan Ning*[1], *Norrina Allen*[1], *John Wilkins*[1], *Kiang Liu*[1], *Donald Lloyd-Jones*[1] *and Lihui Zhao*[1]. [1]Northwestern University Feinberg School of Medicine [2]Washington University School of Medicine in St. Louis [3]Northwestern University [4]Weill Cornell Medicine [5]University of California, Santa Barbara

16:25(EDT) Statistical Disaggregation — a Monte Carlo approach under Constraints
Shenggang Hu, ♦*Hongsheng Dai and Fanlin Meng*. University of Essex

16:50(EDT) Multimodal Neuroimaging Data Integration and Pathway Analysis
♦*Yi Zhao*[1], *Lexin Li*[2] *and Brian Caffo*[3]. [1]Indiana University School of Medicine [2]University of California Berkeley [3]Johns Hopkins Bloomberg School of Public Health

17:15(EDT) Disclosure control for microdata: a mixture modeling approach
♦*Bei Jiang*[1], *Adrian Raftery*[2], *Russel Steele*[3] *and Naisyin Wang*[4]. [1]University of Alberta [2]University of Washington [3]McGill University [4]University of Michigan

17:40(EDT) Floor Discussion.

Session 51: Statistical Methods for Single-Cell Data
Organizer: Yun Li.
Chair: Yun Li.

16:00(EDT) SnapHiC: a computational pipeline to identify chromatin loops from single cell Hi-C data
Ming Hu. Cleveland Clinic

16:25(EDT) Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer
Chi-Yun Wu and ♦*Nancy Zhang*. University of Pennsylvania

16:50(EDT) Jointly Defining Cell States from Single-Cell Multi-Omic and Spatial Transcriptomic Datasets

Joshua Welch. University of Michigan

17:15(EDT) TWO-SIGMA-G: A New Competitive Gene Set Testing Framework for scRNA-seq Data

♦*Eric Van Buren*[1]*, Ming Hu*[2]*, Liang Cheng*[3]*, John Wrobel*[3]*, Kirk Wilhelmsen*[3]*, Lishan Su*[3]*, Yun Li*[3] *and Di Wu*[3]. [1]Harvard University [2]Cleveland Clinic [3]University of North Carolina at Chapel Hill

17:40(EDT) Floor Discussion.

Session 52: New Challenges in Functional Data Analysis

Organizer: GUANQUN CAO.
Chair: Guanqun Cao.

16:00(EDT) A reproducing kernel Hilbert space framework for functional classification

♦*Peijun Sang*[1]*, Adam Kashlak*[2] *and Linglong Kong*[2]. [1]University of Waterloo [2]University of Alberta

16:25(EDT) Functional L-Optimality Subsampling for Massive Data

Hua Liu[1]*, Jinhong You*[1] *and* ♦*Jiguo Cao*[2]. [1]Shanghai University of Finance and Economics [2]Simon Fraser University

16:50(EDT) Nonparametric Estimation of Repeated Densities with Heterogeneous Sample Sizes

Jiaming Qiu, ♦*Xiongtao Dai and Zhengyuan Zhu*. Iowa State University

17:15(EDT) Optimal Imperfect Classification for Functional Data

♦*Shuoyang Wang*[1]*, Zuofeng Shang*[2]*, Guanqun Cao*[1] *and Jun Liu*[3]. [1]Auburn University [2]New jersey Institute of Technology [3]Harvard University

17:40(EDT) Floor Discussion.

Session 53: Master Protocols - Theories, Applications, and When to Use It

Organizer: Ling Wang.
Chair: Satrajit Roychoudhury.

16:00(EDT) FDA Review Perspectives on Master Protocols in Oncology

Erik Bloomquist. FDA

16:25(EDT) Methodological Challenges in Collaborative Platform Trials

♦*Martin Posch, Marta Bofill Roig and Franz Konig*. Medical University of Vienna

16:50(EDT) Simultaneous False-Regulatory Error Rates in Master protocols with Shared Control: False Discovery Rate Perspective

Jingjing Ye. BeiGene

17:15(EDT) Practical Considerations of Implementing Master Protocols for non-oncology studies in Industry

♦*Ling Wang, Pranab Ghosh and Satrajit Roychoudhury*. Pfizer

17:40(EDT) Floor Discussion.

Virtual Poster & Mixer

Organizer: Fang Jin.
Chair: Fang Jin.

18:05(EDT) Ambiguity Resolution to Enhance BERT Question-Answering

♦*Muzhe Guo*[1]*, Muhao Guo*[2]*, Edward Dougherty*[3] *and Fang Jin*[1]. [1]the George Washington University [2]Arizona State University [3]Roger Williams University

18:07(EDT) An Interactive Knowledge Graph Based Platform for COVID-19Clinical Research

♦*Juntao Su*[1]*, Edward Dougherty*[2]*, Shuang Jiang*[1] *and Fang Jin*[1]. [1]George Washington University [2]Roger Williams University

18:09(EDT) Machine-Understandable Saliency Estimation In Natural Language Processing

♦*Zhou Yang and Shunyan Luo*. George Washington University

18:11(EDT) Optimal Treatment Decision Rules in Precision Medicine based on Outcome Trajectories and Biosignatures

♦*Lanqiu Yao and Thaddeus Tarpey*. NYU School of Medicine

18:13(EDT) Machine- Understandable Saliency Estimation In Natural Language Processing

♦*SHUNYAN LUO, Zhou Yang and Fang Jin*. George Washington University

18:15(EDT) LRPT: A deep learning-based dynamic framework to improve resistance prediction on longitudinal bacterial samples via Transfer Learning

♦*Yiqing Wang*[1]*, Shidan Wang*[2]*, Jiwoong Kim*[2]*, Sen Yang*[1]*, Guanghua Xiao*[2] *and Xiaowei Zhan*[2]. [1]Southern Methodist University [2]Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America

18:17(EDT) MB-SupCon: An Integrative Modeling Framework to Improve Microbiome-based predictive models via Supervised Contrastive Learning

♦*Sen Yang*[1]*, Shidan Wang*[2]*, Yiqing Wang*[1]*, Jiwoong Kim*[2]*, Dajiang Liu*[3]*, Guanghua Xiao*[2] *and Xiaowei Zhan*[2]. [1]Department of Statistical Science, Southern Methodist University [2]Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwestern Medical Center [3]Department of Public Health Sciences, Pennsylvania State University

18:19(EDT) Reluctant Interaction Modeling in GLM

♦*Hai Lu and Guo Yu*. University of California, Santa Barbara

18:21(EDT) Sensitivity Analysis of Causal Treatment Effect Estimation for Clustered Observational Data with Unmeasured Confounding

♦*Yang Ou, Lu Tang and Chung-Chou Chang*. University of Pittsburgh

18:23(EDT) On the Bayesian Multiple Index Additive Models

♦*Zhengkang Liang and Zhigen Zhao*. Temple University

18:25(EDT) Mixture of Shape-on-Scalar Regression Models: Going Beyond Prealigned Non-Euclidean Responses
♦*Yuanyao Tan*[1] *and Chao Huang*[2]. [1]Florida State University [2]chuang7@fsu.edu

18:27(EDT) An Eigenmodel for Dynamic Multilayer Networks
♦*Joshua Loyal and Yuguo Chen*. University of Illinois at Urbana-Champaign

18:29(EDT) A Sparse Projection Regression Framework for Generalized Eigenvalue Problems
♦*Dylan Molho*[1]*, Yuying Xie*[1] *and Qiang Sun*[2]. [1]Michigan State University [2]University of Toronto

18:31(EDT) Statistical Learning in Preclinical Drug Proarrhythmic Assessment
♦*Nan Miles Xi*[1]*, Yu-Yi Hsu*[2]*, Qianyu Dang*[2] *and Dalong Patrick Huang*[2]. [1]Loyola University Chicago [2]FDA

18:33(EDT) Artificial intelligence-driven radiogenomic analysis framework with mediation analysis for identifying prognostic radiogenomic biomarkers in breast cancer
♦*Qian Liu and Pingzhao Hu*. University of Manitoba

18:35(EDT) A Novel Model Checking Approach for Dose-Response Relationships
♦*Yu Xia*[1]*, Xinmin Li*[2]*, Hua Liang*[1] *and Shunyao Wu*[2]. [1]George Washington University [2]Qingdao University

18:37(EDT) Semiparametric marginal regression analysis of clustered multistate process data
♦*Wenxian Zhou*[1]*, Giorgos Bakoyannis*[1]*, Ying Zhang*[2] *and Constantin T. Yiannoutsos*[1]. [1]Indiana University [2]University of Nebraska Medical Center

18:39(EDT) A Bayesian Approach to Variable Selection in Binary Quantile Regression
♦*Mai Dao*[1]*, Min Wang*[2] *and Souparno Ghosh*[3]. [1]Texas Tech University [2]University of Texas at San Antonio [3]University of Nebraska-Lincoln

18:41(EDT) A Bayesian Approach for Joint Estimation for Sparse Canonical Correlation and Graphical Models
♦*Siddhesh Kulkarni, Jeremy Gaskins and Subhadip Pal*. University of Louisville

18:43(EDT) Statistical analyses of housekeeping genes' methylation patterns in breast cancer
Shuying Sun. Texas State University

18:45(EDT) Application of machine learning methods in clinical trial design with response?adaptive randomization
♦*Yizhuo Wang*[1]*, Xuelin Huang*[1]*, Ziyi Li*[1] *and Bing Carter*[2]. [1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center [2]Department of Leukemia, The University of Texas MD Anderson Cancer Center

18:47(EDT) Comparison of data-driven subgroup identification methods for binary and time-to-event outcome
♦*Yujie Zhao*[1]*, Ahrim Youn*[2]*, Yanping Chen*[2] *and Casey Xu*[2]. [1]The University of Texas, MD Anderson Cancer Center [2]Bristol Myers Squibb

18:49(EDT) Robust Inference for Linear Models under Huber's Contamination Model
♦*Peiliang Zhang*[1]*, Wen-Xin Zhou*[2] *and Zhao Ren*[1]. [1]University of Pittsburgh [2]University of California, San Diego

18:51(EDT) Application of multiple criteria decision analysis (MCDA) in early phase drug development
♦*Weibin Zhong*[1]*, Alan Wu*[2]*, Casey Xu*[2]*, Ahrim Youn*[2] *and Jun Zhang*[2]. [1]Bristol Myers Squibb and George Mason University [2]Bristol Myers Squibb

18:53(EDT) Applying Group Sequential Design with Multiple Comparison Consideration in Confirmatory Clinical Trial
♦*Shuyu Chu and Min Chen*. BMS

18:55(EDT) An Application of Predictive Modelling in Supporting Fedratinib Regulatory Filing
♦*Jun Zhang, Alan Wu and Ivan Chan*. BMS

18:57(EDT) New Class of Distortion Risk Measures and Their Estimation
Xiwen Wang. Central Michigan University

18:59(EDT) On Construction and Estimation of Stationary Mixture Transition Distribution Models
♦*Xiaotian Zheng, Athanasios Kottas and Bruno Sansó*. University of California Santa Cruz

19:01(EDT) Floor Discussion.

Sep. 14 10:20-noon (EDT)

Session 54: Nonconvex Optimization in Statistics

Organizer: Anru Zhang.
Chair: Anru Zhang.

10:20(EDT) Linear Polytree Structural Equation Models: Structural Learning and Inverse Correlation Estimation
Xingmei Lou[1]*, Yu Hu*[2] *and* ♦*Xiaodong Li*[1]. [1]UC Davis [2]Hong Kong University of Science and Technology

10:45(EDT) Compositional Optimization under Misspecification
Ethan Fang. Penn State

11:10(EDT) Sample complexity of Q-learning: sharper analysis and variance reduction
Gen Li[1]*,* ♦*Yuting Wei*[2]*, Yuejie Chi*[3]*, Yuantao Gu*[4] *and Yuxin Chen*[1]. [1]Princeton University [2]University of Pennsylvania [3]CMU [4]Tsinghua University

11:35(EDT) Asymptotic Analysis of Accelerated Stochastic Gradient Descent
Shang Wu. Fudan University

12:00(EDT) Floor Discussion.

Sep. 14 10:20-noon (EDT)

Session 55: Transforming Industries with Advanced Data Science Solutions

Organizer: Sijian Wang.
Chair: Sijian Wang.

10:20(EDT) Lessons learned from AlphaGo Zero to AlphaFold 2
Haoda Fu. Eli Lilly and Company

10:45(EDT) A few things about product data scientist
♦*Jie Cheng and Yuan Jin*. Google

11:10(EDT) A Time To Event Framework For Multi-touch Attribution
Dinah Shender[1]*, Ali Nasiri Amini*[1]*, Xinlong Bao*[1]*, Mert Dikmen*[1]*, Amy Richardson*[2] *and* ♦*Jing Wang*[1]. [1]Google [2]At Google when this work was done

11:35(EDT) Floor Discussion.

Session 56: Statistical Considerations for Data Integration and Data Privacy
Organizer: DI Shu.
Chair: Di Shu.

10:20(EDT) Integrating Information from Existing Risk Prediction Models with No Model Details
♦*Peisong Han, Jeremy Taylor and Bhramar Mukherjee*. University of Michigan

10:45(EDT) Privacy-protecting Cox Proportional Hazards Regression for Distributed Data Networks Using Summary-level Information
Dongdong Li[1]*, Wenbin Lu*[2]*, Di Shu*[3]*, Sengwee Toh*[1] *and* ♦*Rui Wang*[1]. [1]Harvard Pilgrim Health Care Institute [2]North Carolina State University [3]University of Pennsylvania

11:10(EDT) Gaussian Differential Privacy
Weijie Su. University of Pennsylvania

11:35(EDT) Distributed algorithms for mixed effects models
♦*Yong Chen and Chongliang Luo*[1]. [1]Washington University at St Louis

12:00(EDT) Floor Discussion.

Session 57: Recent developments in survival and recurrent event data analysis
Organizer: Ming-Yueh Huang.
Chair: Ming-Yueh Huang.

10:20(EDT) Testing the proportional hazard assumption under monotonicity constraints
Huan Chen and ♦*Chuan-Fa Tang*. University of Texas at Dallas

10:45(EDT) Weighted least-squares estimation for semiparametric accelerated failure time model with regularization
Ying Chen, Chuan-Fa Tang, ♦*Sy Han Chiou and Min Chen*. University of Texas at Dallas

11:10(EDT) Estimation and Model Checking for General Semiparametric Recurrent Event Models with Informative Censoring
Hung-Chi Ho. China Medical University and Hospital, Taiwan

11:35(EDT) Unification of semicompeting risks analysis through causal mediation modeling
♦*Jin-Chang Yu and Yen-Tsung Huang*. Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

12:00(EDT) Floor Discussion.

Session 58: Recent developments in statistical network data analysis
Organizer: Binyan Jiang.
Chair: Binyan Jiang.

10:20(EDT) Community Detection on Mixture Multi-layer Networks via Regularized Tensor Decomposition
Bing-Yi JING[1]*, Ting LI*[2]*, Zhongyuan LYU*[1] *and* ♦*Dong XIA*[1]. [1]HKUST [2]HK PolyU

10:45(EDT) Network Structure Inference from Grouped Observations
♦*Yunpeng Zhao*[1]*, Peter Bickel*[2] *and Charles Weko*[3]. [1]Arizona State University [2]University of California, Berkeley [3]US Army

11:10(EDT) Two-sample inference for network moments
♦*Yuan Zhang*[1]*, Dong Xia*[2]*, Qiong Wu*[3] *and Shuo Chen*[3]. [1]Ohio State University [2]Hong Kong University of Science and Technology [3]University of Maryland

11:35(EDT) Autoregressive Networks
♦*Binyan Jiang*[1]*, Jialiang Li*[2] *and Qiwei Yao*[3]. [1]The Hong Kong Polytechnic University [2]National University of Singapore [3]London School of Economics

12:00(EDT) Floor Discussion.

Session 59: Recent development on spatial statistics
Organizer: Wenlin Dai.
Chair: Wenlin Dai.

10:20(EDT) Bayesian space-time gap filling for inference on extreme hotspots: an application to Red Sea surface temperatures
♦*Daniela Castro-Camilo*[1]*, Linda Mhalla*[2] *and Thomas Opitz*[3]. [1]University of Glasgow [2]HEC Lausanne [3]Biostatistics and Spatial Processes, INRAE, Avignon, France

10:45(EDT) Global Wind Modeling with Transformed Gaussian Processes
Jaehong Jeong. Hanyang University

11:10(EDT) Modeling Spatial Data with Cauchy Convolution Processes
♦*Pavel Krupskiy*[1] *and Raphael Huser*[2]. [1]University of Melbourne [2]King Abdullah University of Science and Technology

11:35(EDT) Regime based Precipitation Modeling
Carolina Euan. Lancaster University

12:00(EDT) Floor Discussion.

Session 60: Recent advances in statistical methods for modern degradation data
Organizer: Zhisheng Ye.
Chair: Zhisheng Ye.

10:20(EDT) Planning Accelerated Degradation Tests with Two Stress Variables
♦*Guanqi Fang*[1]*, Rong Pan*[2] *and John Stufken*[3]. [1]Zhejiang Gongshang University [2]Arizona State University [3]University of North Carolina-Greensboro

10:45(EDT) Accelerated degradation tests with inspection effects
Xiujie Zhao. Tianjin University

11:10(EDT) A generalized Wiener process with dependent degradation rate and volatility and time-varying mean-to-variance ratio
Shirong Zhou[1]*, Yincai Tang*[1] *and* ♦*Ancha Xu*[2]. [1]East China Normal University [2]Zhejiang Gongshang University

11:35(EDT) Pairwise model discrimination with applications in lifetime distributions and degradation processes
♦*Piao Chen*[1]*, Zhisheng Ye*[2] *and Xun Xiao*[3]. [1]TU Delft [2]National University of Singapore [3]University of Otago

12:00(EDT) Floor Discussion.

Session 61: Recent Developments in Meta-Analysis
Organizer: Qixuan Chen.
Chair: Yajuan Si.

10:20(EDT) A Bayesian model for combining standardized mean differences and odds ratios in the same meta-analysis
Yaqi Jing[1]*, Mohammad Hassan Murad*[2] *and* ♦*Lifeng Lin*[1]. [1]Florida State University [2]Mayo Clinic

10:45(EDT) Bivariate hierarchical Bayesian model for combining summary measures and their uncertainties from multiple sources
Yujing Yao[1]*, R. Todd Ogden*[1]*, Chubing Zeng*[2] *and* ♦*Qixuan Chen*[1]. [1]Columbia University [2]University of Southern California

11:10(EDT) Quantifying Replicability of Multiple Studies in a Meta-Analysis
♦*Mengli Xiao*[1]*, Haitao Chu*[1]*, James Hodges*[1] *and Lifeng Lin*[2]. [1]University of Minnesota [2]Florida State University

11:35(EDT) Bayesian inference for asymptomatic Covid-19 infection rates
Dexter Cahoy[1] *and* ♦*Joseph Sedransk*[2]. [1]University of Houston [2]University of Maryland

12:00(EDT) Floor Discussion.

Session 62: Modern streaming Data Analysis: Sequential Tests
Organizer: Ruizhi Zhang.
Chair: Yajun Mei.

10:20(EDT) On Sequential Test with Optimal Observation Strategy for Detection of Change in Sensor Networks
♦*Dong Han*[1]*, Fugee Tsung*[2] *and Lei Qiao*[1]. [1]Shanghai Jiao Tong University [2]Hong Kong University of Science and Technology

10:45(EDT) Adversarially Robust Sequential Hypothesis Testing
Shuchen Cao[1]*,* ♦*Ruizhi Zhang*[1] *and Shaofeng Zou*[2]. [1]University of Nebraska-Lincoln [2]University at Buffalo, The State University of New York

11:10(EDT) A multistage test for high-dimensional signal recovery
Yiming Xing and ♦*Georgios Fellouris*. University of Illinois, Urbana-Champaign

11:35(EDT) Asymptotically optimal sequential FDR and pFDR control
Jay Bartroff. University of Southern California

12:00(EDT) Floor Discussion.

Session 63: Modern Statistical Methods for Complex Data Objects Analysis
Organizer: Lily Wang.
Chair: Shan Yu.

10:20(EDT) Unified Principal Component Analysis for Sparse and Dense Functional Data under Spatial Dependency
Yehua Li. University of California at Riverside

10:45(EDT) Spline smoothing of 3D geometric data
♦*Xinyi Li*[1]*, Shan Yu*[2]*, Yueying Wang*[3]*, Guannan Wang*[4] *and Lily Wang*[5]. [1]Clemson Univeristy [2]University of Virginia [3]Columbia University [4]College of William and Mary [5]George Mason University

11:10(EDT) Sequential Change-point Detection for High-Dimensional and non-Euclidean Data
♦*Lynna Chu*[1] *and Hao Chen*[2]. [1]Iowa State University [2]University of California, Davis

11:35(EDT) Broadcasted Nonparametric Tensor Regression
Ya Zhou[1]*,* ♦*Raymond K. W. Wong*[2] *and Kejun He*[3]. [1]Renmin University of China and Texas A&M University [2]Texas A&M University [3]Renmin University of China

12:00(EDT) Floor Discussion.

Session 64: On inference for time series models
Organizer: Qianqian Zhu.
Chair: Qianqian Zhu.

10:20(EDT) Location-adaptive change-point testing for time series
Linlin Dai and ♦*Rui She*. Southwestern University of Finance and Economics

10:45(EDT) Statistical inferences for threshold GARCH model based on the intraday high frequency data
Xingfa Zhang. Guangzhou University

11:10(EDT) Bootstrapping on robust goodness-of-fit test for GARCH models
♦*Xuqin Wang and Muyi Li*. Xiamen University

11:35(EDT) Smooth transition moving average models: estimation, testing, and computation
♦*Xinyu Zhang and Dong Li*. Tsinghua University

12:00(EDT) Floor Discussion.

Session 65: recent advances in nonparametric and robust estimation and inference for complex data
Organizer: Shujie Ma.
Chair: Guanyu Hu.

10:20(EDT) New Regression Model: Modal Regression
♦*Weixin Yao*[1]*, Longhai Li*[2] *and Sijia Xiang*[3]. [1]University of California, Riverside [2]University of Saskatchewan [3]Zhejiang University of Finance & Economics

10:45(EDT) Big spatial data learning: a parallel solution
Shan Yu[1]*, Guannan Wang*[2] *and* ♦*Lily Wang*[3]. [1]University of Virginia [2]College of William and Mary [3]George Mason University

11:10(EDT) Spatial Homogeneity Regression: A Bayesian Nonparametric Recourse
Guanyu Hu. University of Missouri

11:35(EDT) BEAUTY powered BEAST
♦*Kai Zhang*[1]*, Zhigen Zhao*[2] *and Wen Zhou*[3]. [1]UNC Chapel Hill [2]Temple University [3]Colorado State University

12:00(EDT) Floor Discussion.

Session 66: Novel Methods for Statistical Analysis of Complex Data
Organizer: Hua He.
Chair: Hua He.

10:20(EDT) Semiparametric Regression Models for Between- and Within-subject Attributes: Asymptotic Efficiency and Applications to High-Dimensional Data
Xin Tu. Division of Biostat and Bioinfo, UCSD

10:45(EDT) fMRI data classification based on the nonparametric bayesian graph model
Chong Wan and ♦*Guoxin Zuo*. Central China Normal University

11:10(EDT) Diagnosis of thyroid nodules for ultrasonographic characteristics indicative of malignancy using random forest
Dan Chen[1], ♦*Jun Hu*[2], *Mei Zhu*[3], *Niansheng Tang*[1], *Yang Yang*[3] *and Yuran Feng*[3]. [1]Yunnan University [2]Yunnan Agricultural University [3]The First Affiliated Hospital of Kunming Medical University

11:35(EDT) A class of weighted estimating equations for additive hazard models with covariates missing at random
PENG YE. School of Statistics, University of International Business and Economics, Beijing 100029, China

12:00(EDT) Floor Discussion.

Session 67: Advances in Models and Methods for Complex Spatial Processes
Organizer: Yiping Hong.
Chair: Yiping Hong.

10:20(EDT) The frequency and severity of crop damage by wildlife in rural Beijing, China
♦*Liang Fang*[1], *Yiping Hong*[2], *Zaiying Zhou*[3] *and Wenhui Chen*[1]. [1]Beijing Forestry University [2]King Abdullah University of Science and Technology [3]Tsinghua University

10:45(EDT) Copula-based Multiple Indicator Kriging for non-Gaussian Random Fields
Gaurav Agarwal. Statistics at Lancaster University, UK

11:10(EDT) Efficient Calibration of Numerical Model Output Using Hierarchical Dynamic Models
♦*Yewen Chen*[1], *Xiaohui Chang*[2] *and Hui Huang*[1]. [1]Sun Yat-sen university [2]Oregon State University

11:35(EDT) A Nonstationary Spatio-Temporal Autologistic Regression Model
♦*Yiping Hong, Marc G. Genton and Ying Sun*. King Abdullah University of Science and Technology

12:00(EDT) Floor Discussion.

Session 68: Recent development in complex dependent data analysis
Organizer: Guodong Li.
Chair: Guodong Li.

10:20(EDT) Spiked Eigenvalues of High-dimensional Sample Autocovariance Matrices: CLT and Applications
Daning Bi[1], *Xiao Han*[2], *Adam Nie*[1] *and* ♦*Yanrong Yang*[1]. [1]The Australian National University [2]University of Science and Technology of China

10:45(EDT) Quantile index regression
♦*Yingying Zhang*[1], *Jingyu Zhao*[2], *Guodong Li*[2] *and Chil-Ling Tsai*[3]. [1]East China Normal University [2]University of Hong Kong [3]University of California at Davis

11:10(EDT) SARMA: A Novel Computationally Scalable High-Dimensional Vector Autoregressive Moving Average Model
♦*Feiqing Huang*[1], *Yao Zheng*[2], *Di Wang*[3] *and Guodong Li*[1]. [1]University of Hong Kong [2]University of Connecticut [3]University of Chicago

11:35(EDT) Robust estimation of high-dimensional vector autoregressive models
♦*Di Wang and Ruey S. Tsay*. University of Chicago

12:00(EDT) Floor Discussion.

Panel discussion: Statistics and Data Science Partnerships and Collaborations across Sectors

Organizer: Kelly Zou, Viatris.
Chair: Kelly Zou, Viatris.

Panelist	Affiliation
Aniketh Talwai	Medidata, a Dassault Systèmes Company
Kimberly Sellers	Georgetown University
Kelly Zou	Viatris

Sep. 14 2-3:40pm(EDT)

Session 69: Novel statistical models of -omic data analysis
Organizer: Xiaoyu Song.
Chair: Xiaoyu Song.

14:00(EDT) DiSNEP: a Disease-Specific gene Network Enhancement to improve Prioritizing candidate disease genes
Peifeng Ruan[1] *and* ♦*Shuang Wang*[2]. [1]Yale University [2]Columbia University

14:25(EDT) EPIC: inferring relevant cell types for complex traits by integrating genome-wide association studies and single-cell RNA sequencing
Rujin Wang, Danyu Lin and ♦*Yuchao Jiang*. University of North Carolina at Chapel Hill

14:50(EDT) Robust estimation of cell type fractions from bulk data via ensemble of various deconvolution approaches
♦*Jiebiao Wang, Manqi Cai and Wei Chen*. University of Pittsburgh

15:15(EDT) Bayesian Genome-wide TWAS method integrating both cis- and trans- eQTL with GWAS summary statistics
Justin Luningham[1], *Junyu Chen*[2], *Shizhen Tang*[2], *Philip De Jager*[3], *David Bennett*[4], *Aron Buchman*[4] *and* ♦*Jingjing Yang*[2]. [1]Georgia State University [2]Emory University [3]Columbia University Irving Medical Center [4]Rush University Medical Center

15:40(EDT) Floor Discussion.

Session 70: Statistical methods for spatial genomics and transcriptomics
Organizer: Yun Li.
Chair: Yun Li.

14:00(EDT) Statistical methods for spatial genomics and transcriptomics
♦*Mingyao Li, Jian Hu, Kyle Coleman and Amelia Schroeder*. University of Pennsylvania

14:25(EDT) Learning fine-resolution Hi-C contact maps
Wenxiu Ma. University of California Riverside

14:50(EDT) Spatial Transcriptomics Analysis
Guo-Cheng Yuan. Icahn School of Medicine at Mount Sinai

15:15(EDT) MUNIn: A statistical framework for identifying long-range chromatin interactions from multiple samples
♦*Yuchen Yang*[1]*, Weifang Liu*[1]*, Yun Li*[1] *and Ming Hu*[2]. [1]University of North Carolina at Chapel Hill [2]Cleveland Clinic Foundation

15:40(EDT) Floor Discussion.

Session 71: Structural Inference of Time Series Data
Organizer: Weichi Wu.
Chair: Weichi Wu.

14:00(EDT) Multiple change point detection under serial dependence
♦*Haeran Cho*[1] *and Piotr Fryzlewicz*[2]. [1]University of Bristol [2]London School of Economics

14:25(EDT) Modeling the COVID-19 infection trajectory: a piecewise linear quantile trend model
♦*Feiyu Jiang*[1]*, Zifeng Zhao*[2] *and Xiaofeng Shao*[3]. [1]Fudan University [2]University of Notre Dame [3]University of Illinois at Urbana Champaign

14:50(EDT) Testing long-range dependence for locally stationary time series nonparametric regression
♦*Lujia Bai and Weichi Wu*. Tsinghua University

15:15(EDT) A Composite Likelihood-based Approach for Change-point Detection in Spatio-temporal Process
Ting Fung Ma[1]*, Wai Leong Ng*[2]*,* ♦*Chun Yip Yau*[3] *and Zifeng Zhao*[4]. [1]University of Wisconsin, Madison [2]Hang Seng University of Hong Kong [3]Chinese University of Hong Kong [4]University of Notre Dame

15:40(EDT) Floor Discussion.

Session 72: Advances in Forensic Statistics
Organizer: Martin Slawski.
Chair: Martin Slawski.

14:00(EDT) Homogeneity test for ordinal ROC regression and application to facial recognition
♦*Ngoc-Ty Nguyen and Larry Tang*. National Center of Forensic Science, University of Central Florida

14:25(EDT) The Effect of Latent Structures on Forensic Values of Evidence
Dylan Borchet, Andrew Simpson, Semhar Michael and ♦*Christopher Saunders*. South Dakota State University

14:50(EDT) Constructing Coherent Score-Based Likelihood Ratios that Account for Rarity
♦*Danica Ommen*[1] *and Nate Garton*[2]. [1]Iowa State University/CSAFE [2]Commonwealth Computer Research, Inc.

15:15(EDT) Approaches to Likelihood Ratio Estimation for Forensic Evidence Interpretation
♦*He Qi and Martin Slawski*. George Mason University

15:40(EDT) Floor Discussion.

Session 73: Advances in selective and simultaneous inference
Organizer: Shih-Kang Chao.
Chair: Shih-Kang Chao.

14:00(EDT) Post-selection inference: some answers and some questions
Arun KUCHIBHOTLA. Carnegie Mellon University

14:25(EDT) Valid Inference After Hierarchical Clustering
♦*Lucy Gao*[1]*, Jacob Bien*[2] *and Daniela Witten*[3]. [1]University of Waterloo [2]University of Southern California [3]University of Washington

14:50(EDT) Selective peak effect size inference
♦*Samuel Davenport*[1] *and Thomas Nichols*[2]. [1]University of California, San Diego [2]University of Oxford

15:15(EDT) Floor Discussion.

Session 74: Advanced statistical inference for complex data structures
Organizer: Tianxi Li.
Chair: Tianxi Li.

14:00(EDT) Hypothesis tests for block Markov chains
♦*Can Le and Russell Okino*. University of California, Davis

14:25(EDT) Change Point Detection in Network Sequences
♦*Sharmodeep Bhattacharyya*[1]*, Shirshendu Chatterjee*[2]*, Shyamal De*[3] *and Soumendu Mukherjee*[3]. [1]Oregon State University [2]City University of New York [3]Indian Statistical Institute, Kolkata

14:50(EDT) Partial Recovery for Top-k Ranking: Optimality of MLE and Sub-Optimality of Spectral Method
Anderson Ye Zhang. University of Pennsylvania

15:15(EDT) Floor Discussion.

Session 75: Exploratory Functional Data Analysis
Organizer: Ying Sun.
Chair: Ying Sun.

14:00(EDT) Functional outlier detection and taxonomy by sequential transformations
♦*Wenlin Dai*[1]*, Tomas Mrkvicka*[2]*, Ying Sun*[3] *and Marc G. Genton*[3]. [1]Renmin University of China [2]University of South Bohemia [3]King Abdullah University of Science and Technology

14:25(EDT) Covariance function visualization using functional data analysis
♦*Huang Huang, Ying Sun and Marc Genton*. King Abdullah University of Science and Technology

14:50(EDT) Sparse Functional Boxplots for Multivariate Curves
Zhuo Qu and ♦*Marc G. Genton*. KAUST

15:15(EDT) Exploratory Functional Data Analysis
Hanlin Shang. Macquarie University

15:40(EDT) Floor Discussion.

Session 76: New Frontiers in Precision Medicine
Organizer: Lily Wang.
Chair: Xinyi Li.

14:00(EDT) Learning Individualized Treatment Rules for Multiple-Domain Latent Outcomes
Yuan Chen[1], *Yuanjia Wang*[2] *and* ♦*Donglin Zeng*[3]. [1]yc3281@cumc.columbia.edu [2]yw2016@cumc.columbia.edu [3]dzeng@email.unc.edu

14:25(EDT) Estimation and inference on individualized treatment rule in observational data.
Yingqi Zhao. Fred Hutchinson Cancer Research Center

14:50(EDT) Experimental Design and Causal Inference Methods For Micro-Randomized Trials: A Framework for Developing Mobile Health Interventions
♦*Tianchen Qian*[1], *Predrag Klasnja*[2] *and Susan Murphy*[3]. [1]University of California, Irvine [2]University of Michigan [3]Harvard University

15:15(EDT) Causal Inference on Non-linear Spaces: Distribution Functions and Beyond
Zhenhua Lin[1], ♦*Dehan Kong*[2] *and Linbo Wang*[2]. [1]National University of Singapore [2]University of Toronto

15:40(EDT) Floor Discussion.

Session 77: Efficient estimation methods for clinical trials
Organizer: Zhiwei Zhang.
Chair: Chenguang Wang.

14:00(EDT) everaging auxiliary covariates to improve efficiency of inferences: a general framework and practical considerations
♦*Min Zhang*[1] *and Baqun Zhang*[2]. [1]University of Michigan [2]Shanghai University of Finance and Economics

14:25(EDT) Covariate adjustment in randomized studies with ordinal and time-to-event endpoints
Ivan Diaz. Weill Cornell Medicine

14:50(EDT) Model-Robust Inference for Clinical Trials that Improve Precision by Stratified Randomization and Covariate Adjustment
♦*Bingkai Wang*[1], *Ryoko Susukida*[2], *Ramin Mojtabai*[2], *Masoumeh Amin-Esmaeili*[2] *and Michael Rosenblum*[3]. [1]The statistics and data science department of the Wharton School, University of Pennsylvania [2]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health [3]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

15:15(EDT) Data-Adaptive Efficient Estimation Strategies for Biomarker Studies Embedded in Randomized Trials
Wei Zhang[1], ♦*Zhiwei Zhang*[2], *James Troendle*[2] *and Aiyi Liu*[2]. [1]Chinese Academy of Sciences [2]National Institutes of Health

15:40(EDT) Floor Discussion.

Session 78: Recent developments in regression methods for bio-medical studies
Organizer: Ao Yuan.
Chair: Ao Yuan.

14:00(EDT) Multicategory Outcome Weighted Learning for Dynamic Treatment Regimes
♦*Wenqing He and Junwei Shen*. University of Western Ontario

14:25(EDT) A Semiparametric Isotonic Regression Model for Skewed Distributions
♦*Chenguang Wang*[1], *Ao Yuan*[2], *Leslie Cope*[1] *and Jing Qin*[3]. [1]Johns Hopkins University [2]Georgetown University [3]National Institute of Allergy and Infectious Diseases

14:50(EDT) Order-Constrained ROC Regression with Application to Facial Recognition
Larry Tang. University of Central Florida

15:15(EDT) Inferring random change point with segmented non-linear mixed effect models
Hongbin Zhang. City University of New York

15:40(EDT) Floor Discussion.

Session 79: Recent Advance of Design of Experiments in Online Experimentation and Personalized Medicine
Organizer: Xinwei Deng.
Chair: Devon Lin.

14:00(EDT) Novelty and Primacy: A Long-Term Estimator for Online Experiments
♦*Sol Sadeghi, Somit Gupta, Stefan Gramatovici, Jiannan Lu, Hao Ai and Ruhan Zhang*. Microsoft

14:25(EDT) Robust sequential design for piecewise-stationary multi-armed bandit problem in the presence of outliers
Yaping Wang[1], *Zhicheng Peng*[1], *Riquan Zhang*[1] *and* ♦*Qian Xiao*[2]. [1]East China Normal University [2]University of Georgia

14:50(EDT) Min-Max Optimal Design of Two-Armed Trials with Side Information
♦*Qiong Zhang, Amin Khademi and Yongjia Song*. Clemson University

15:15(EDT) Recent Advance in DoE Driven by New Applications and/or Algorithms
Lulu Kang. Illinoise Institue of Techonogy

15:40(EDT) Floor Discussion.

Session 80: Statistical Modeling for the COVID-19 Pandemic
Organizer: Yuedong Wang.
Chair: Yuedong Wang.

14:00(EDT) A Spatiotemporal Epidemiological Prediction Model to Inform County-level COVID-19 Risk in the USA
Yiwang Zhou, Lili Wang and ♦*Peter Song*. University of Michigan

14:25(EDT) Robust estimation of SARS-CoV-2 epidemic in US counties
Hanmo Li and ♦*Mengyang Gu*. University Of California, Santa Barbara

14:50(EDT) SaucIR: a Migration-based Model for Forecasting Confirmed Cases of the COVID-19 Pandemic
Xinyu Wang[1], *Lu Yang*[1], *Hong Zhang*[1], *Zhouwang Yang*[1] *and* ♦*Catherine Liu*[2]. [1]University of Science and Technology of China [2]The Hong Kong Polytechnic University

15:15(EDT) Modelling biological change in COVID-19 patient using nonparametric mixed effect mixture model
Xiaoran Ma. University of California Santa Barbara

15:40(EDT) Floor Discussion.

Session 81: Incorporating RWE in regulatory submission
Organizer: Meijing Wu.
Chair: Hongtao Zhang.

14:00(EDT) Indirect comparisons using real world data: confounding and population adjustment for time-to event endpoints
♦*Jixian Wang, Hongtao Zhang and Ram Tiwari*. BMS

14:25(EDT) Use of external control for single arm registrational trial: a case study in NSCLC
♦*Jianchang Lin and Qing Li*. Takeda

14:50(EDT) Real-World Data in Regulatory Science
Diqiong Xie. Seagen

15:15(EDT) Use of Real-World Evidence (RWE) in HTA Decision Making
Julia Ma. AbbVie

15:40(EDT) Floor Discussion.

Session 82: Flexible and efficient Bayesian methods for complex data modeling
Organizer: Weining Shen.
Chair: Weining Shen.

14:00(EDT) New Directions in Bayesian Shrinkage for Structured Data
Jyotishka Datta. Virginia Tech

14:25(EDT) A Bayesian Selection Model for Correcting and Quantifying the Impact of Outcome Reporting Bias in Multivariate Meta-Analysis
♦*Ray Bai*[1], *Lifeng Lin*[2], *Mary Boland*[3] *and Yong Chen*[3]. [1]University of South Carolina [2]Florida State University [3]University of Pennsylvania

14:50(EDT) An approximate Bayesian approach to covariate dependent graphical modeling
Sutanoy Dasgupta, Prasenjit Ghosh, ♦*Debdeep Pati and Bani Mallick*. Texas A&M University

15:15(EDT) A Phase I-II Basket Trial Design to Optimize Dose-Schedule Regimes Based on Delayed Outcomes
Ruitao Lin. The University of Texas MD Anderson Cancer Center

15:40(EDT) Floor Discussion.

Session 83: New advances in complex biomedical data analysis and the novel applications
Organizer: Yichuan Zhao.
Chair: Yichuan Zhao.

14:00(EDT) Simple method for detecting the effect of exposure mixture
Xinhua Liu and ♦*Zhezhen Jin*. Columbia University

14:25(EDT) Integrative Clustering Analysis with Application in Multi-Source Gene Expression Data
Liuqing Yang[1], *Yunpeng Zhao*[2] *and* ♦*Qing Pan*[1]. [1]George Washington University [2]Arizona State University

14:50(EDT) Peer-to-Peer Masking for Privacy-Preserving Medical Data Sharing
Hanzhi Gao[1], *Dimitrios Melissourgos*[1], *Shigang Chen*[1], *Adam Ding*[2] *and* ♦*Samuel S. Wu*[1]. [1]University of Florida [2]Northeastern University

15:15(EDT) Event-Specific Win Ratios for Inference with Semi-Competing Risks Data
♦*Song Yang*[1], *James Troendle*[2], *Daewoo Pak*[3] *and Eric Leifer*[2]. [1]NIH NHBLI [2]NIH NHLBI [3]Yonsei University, South Korea

15:40(EDT) Floor Discussion.

Sep. 14 4-5:40pm(EDT)

Session 84: Statistics in Imaging
Organizer: Anru Zhang.
Chair: Anru Zhang.

16:00(EDT) Change-point analysis for modern data
♦*Hao Chen*[1], *Shizhe Chen*[1] *and Xinyi Deng*[2]. [1]University of California, Davis [2]Beijing University of Technology

16:25(EDT) Quantification in Colocalization Analysis: Beyond "Red+Green=Yellow"
Shulei Wang. University of Illinois at Urbana-Champaign

16:50(EDT) Methodology for the Comparison of Diffusion Tensor Images Across Cohorts
♦*Gina-Maria Pomann*[1], *Ana-Maria Staicu*[2] *and Sujit Ghosh*[2]. [1]Duke University [2]North Carolina State University

17:15(EDT) Longitudinal ComBat: A Method for Harmonizing Longitudinal Multi-scanner Imaging Data
Joanne Beer, Russell Shinohara and ♦*Kristin Linn*. University of Pennsylvania

17:40(EDT) Floor Discussion.

Session 85: Advances in false discovery rate control methods
Organizer: Tracy Ke.
Chair: Tracy Ke.

16:00(EDT) Data splitting methods for FDR controls
Chenguang Dai[1], *Buyu Lin*[2], *Xing Xin*[3] *and* ♦*Jun Liu*[4]. [1]Citadel LLC [2]Harvard University [3]Virginia Tech [4]Harvard University

16:25(EDT) Inference with approximate co-sufficient sampling
♦*Rina Barber*[1], *Lucas Janson*[2] *and Wanrong Zhu*[1]. [1]University of Chicago [2]Harvard University

16:50(EDT) Multiple testing under dependence and non-sparsity
Hongyuan Cao. FSU

17:15(EDT) Interactive identification of individuals with positive treatment effect while controlling false discoveries
♦*Boyan Duan*[1], *Aaditya Ramdas*[2] *and Larry Wasserman*[2]. [1]Google [2]Carnegie Mellon University

17:40(EDT) Floor Discussion.

Session 86: New classification methods for imaging, network and dynamic treatment
Organizer: Annie Qu.
Chair: Xiwei Tang.

16:00(EDT) Community Detection in Hypergraphs
Yaoming Zhen and ♦*Junhui Wang*. City University of Hong Kong

16:25(EDT) High-order Joint Embedding for Multi-Level Link Prediction
♦*Yubai Yuan and Annie Qu*. University of California, Irvine

16:50(EDT) Solving Infinite Horizon Dynamic Treatment Regimes: A pT Learning Framework
♦*Wenzhuo Zhou*[1], *Ruoqing Zhu*[1] *and Annie Qu*[2]. [1]University of Illinois Urbana Champaign [2]University of California Irvine

17:15(EDT) Dermoscopic Image Classification with Neural Style Transfer
♦*Yutong Li*[1], *Ruoqing Zhu*[1], *Mike Yeh*[1] *and Annie Qu*[2]. [1]University of Illinois at Urbana-Champaign [2]University of California, Irvine

17:40(EDT) Floor Discussion.

Session 87: The Jiann-Ping Hsu Invited Session on Biostatistical and Regulatory Sciences
Organizer: Lili Yu.
Chair: Lili Yu.

16:00(EDT) Statistical Data Analysis in Pharmaceutical Industry
Kao-tai Tsai. BMS

16:25(EDT) Principles of leading with statistical knowledge and a problem-solving approach to innovation
Stephan Ogenstad. Jiann-Ping Hsu College of Public Health, Georgia Southern University

16:50(EDT) Evaluation of SIMEX extrapolation methods in Accelerated failure time models with covariate measurement error
♦*Manyun Liu, Roshni Modi and Lili Yu*. Georgia Southern University

17:15(EDT) Floor Discussion.

Session 88: Survey Analysis and Design Using Multilevel Regression and Post-stratification
Organizer: Qixuan Chen.
Chair: Qixuan Chen.

16:00(EDT) On the Use of Auxiliary Variables in Multilevel Regression and Poststratification
Yajuan Si. University of Michigan, Ann Arbor

16:25(EDT)

16:50(EDT) Survey Design for Multilevel Regression and Post-Stratification
♦*Yutao Liu*[1], *Lauren Kennedy*[2], *Andrew Gelman*[3] *and Qixuan Chen*[3]. [1]Vertex Pharmaceuticals, Inc. [2]Monash University [3]Columbia University

17:15(EDT) Multilevel regression and post-stratification in R
♦*Lauren Kennedy*[1], *Jonah Gabry*[2], *Rohan Alexander*[3], *Dewi Amaliah*[1] *and Mitzi Morris*[2]. [1]Monash University [2]Columbia University [3]University of Toronto

17:40(EDT) Floor Discussion.

Session 89: Massive Data Analysis
Organizer: Ling Zhou.
Chair: Ling Zhou.

16:00(EDT) A Tree-based Federated Learning Approach for Personalized Treatment Effect Estimation from Heterogeneous Data Sources
Xiaoqing Tan, Chung-Chou Chang and ♦*Lu Tang*. University of Pittsburgh

16:25(EDT) Integrative Causal Inference in the Presence of Study Heterogeneity
Ling Zhou[1], ♦*Xu Shi*[2], *Mengtong Hu*[2] *and Peter Song*[2]. [1]Southwestern University of Finance and Econonmics [2]University of Michigan

16:50(EDT) Multivariate Online Regression Analysis with Heterogeneous Streaming Data
♦*Lan Luo*[1] *and Peter Song*[2]. [1]University of Iowa [2]University of Michigan

17:15(EDT) Center-augmented l2-type regularization for subgroup learning
♦*Ling Zhou*[1], *Ye He*[2], *Huazhen Lin*[1] *and Yingcun Xia*[3]. [1]Southwestern University of Finance and Economics [2]University of Electronic Science and Technology of China [3]National University of Singapore

17:40(EDT) Floor Discussion.

Session 90: Modern streaming data analysis: change-point problems and applications
Organizer: Ruizhi Zhang.
Chair: Jie Chen.

16:00(EDT) Detection of Multiple Transient Changes
♦*Michael Baron*[1] *and Sergey Malov*[2]. [1]American University [2]St. Petersburg State University,

16:25(EDT) Equivariant Variance Estimation for Multiple Change-point Model
♦*Ning Hao*[1], *Yue Niu*[1] *and Han Xiao*[2]. [1]University of Arizona [2]Rutgers University

16:50(EDT) Changepoint detection in autocorrelated ordinal categorical time series
Mo Li and ♦*QiQi Lu*. Virginia Commonwealth University

17:15(EDT) Univariate Likelihood Projections and Characterizations of the Multivariate Normal Distribution Applicable to a Multivariate Change Point Detection
Albert Vexler. Professor

17:40(EDT) Floor Discussion.

Session 91: Recent advances in statistical methods for large-scale omics data
Organizer: Yue Wang.
Chair: Yue Wang.

16:00(EDT) A likelihood-based approach for multivariate categorical response regression in high dimensions
♦*Aaron Molstad*[1] *and Adam Rothman*[2]. [1]University of Florida [2]University of Minnesota

16:25(EDT) Estimating gene-to-trait effect with GWAS and eQTL summary statistics using Bayesian Hierarchical Models

♦*ANQI ZHU*[1]*, Nana Matoba*[2]*, Emma Wilson*[3]*, Amanda Tapia*[4]*, Yun Li*[5]*, Joseph Ibrahim*[4]*, Jason Stein*[2] *and Michael Love*[6]. [1]Department of Biostatistics University of North Carolina at Chapel Hill, 23andMe Inc. [2]Neuroscience Center, Department of Genetics, University of North Carolina at Chapel Hill [3]Department of Genetics, University of North Carolina at Chapel Hill [4]Department of Biostatistics, University of North Carolina at Chapel Hill [5]Department of Biostatistics, Department of Genetics and Department of Computer Science, University of North Carolina at Chapel Hill [6]Department of Biostatistics and Department of Genetics, University of North Carolina at Chapel Hill

16:50(EDT) Developing Trans-ethnic Polygenic Risk Scores Using Empirical Bayes and Super Learning Algorithm

Haoyu Zhang. Harvard T.H. Chan School of Public Health

17:15(EDT) De-biased Lasso for Generalized Linear Models with A Diverging Number of Covariates

♦*Lu Xia*[1]*, Bin Nan*[2] *and Yi Li*[3]. [1]University of Washington [2]University of California, Irvine [3]University of Michigan

17:40(EDT) Floor Discussion.

Session 92: Mixtures, two-sample problems and their applications

Organizer: Jingjing Wu.
Chair: Jingjing Wu.

16:00(EDT) Testing Homogeneity in Contaminated Mixture Models

Guanfu Liu[1]*,* ♦*Yuejiao (Cindy) Fu*[2]*, Wenchen Liu*[3] *and Rongji Mu*[4]. [1]Shanghai University of International Business and Economics [2]York University [3]Shanghai Lixin University of Accounting and Finance [4]Shanghai Jiao Tong University

16:25(EDT) Statistical inference for a relaxation index of stochastic dominance under density ratio model

Weiwei Zhuang. University of Science and Technology of China

16:50(EDT) Estimation of B-spline copula

Xiaoling Dou. Waseda University

17:15(EDT) Stochastic Simulation Models for the Spread of Infectious Diseases

Zhiyong Jin, Lin Xue and ♦*Liqun Wang*. University of Manitoba

17:40(EDT) Floor Discussion.

Session 93: Statistical methods for complex data

Organizer: Yuzhen Zhou.
Chair: Yuzhen Zhou.

16:00(EDT) Functional Data Analysis Using Neural Networks

Sneha Jadhav. Wake Forest

16:25(EDT) Bias-corrected estimation in errors-in-variables Logistic regression under case-control study

♦*Pei Geng and Huyen Nguyen*. Illinois State University

16:50(EDT) A unified framework for change point detection in high-dimensional linear models

♦*Abolfazl Safikhani and Yue Bai*. University of Florida

17:15(EDT) Rapid Online Plant Leaf Area Change Detection with High-Throughput Plant Image Data

Yinglun Zhan, Ruizhi Zhang, ♦*Yuzhen Zhou, Vincent Stoerger, Jeremy Hiller, Tala Awada and Yufeng Ge*. University of Nebraska Lincoln

17:40(EDT) Floor Discussion.

Session 94: Frontiers in Semi-Parametric and Non-Parametric Methods

Organizer: Jing Wang.
Chair: Jing Wang.

16:00(EDT) Statistical inference for functional time series: autocovariance function

Chen Zhong and ♦*Lijian Yang*. Tsinghua University

16:25(EDT) Estimation of the Mean Function of Functional Data via Deep Neural Networks

Shuoyang Wang[1]*,* ♦*Guanqun Cao*[1] *and Zuofeng Shang*[2]. [1]Auburn University [2]New Jersey Institute of Technology

16:50(EDT) Two-Step Time Series Modelling

Lijian Yang[1]*,* ♦*Qin Shao*[2]*, Yuanyuan Zhang*[1]*, Hanh Nguyen*[2]*, Rawiyah Alraddadi*[2] *and Rong Liu*[2]. [1]Tsinghua University [2]University of Toledo

17:15(EDT) Free-knot Splines for Generalized Regression Models

♦*Jing Wang and Elena Graetz*. University of Illinois at Chicago

17:40(EDT) Floor Discussion.

Session 95: Statistical Challenges for Single-cell RNA Sequencing Data Analysis

Organizer: Xiaoxiao Sun.
Chair: Xiaoxiao Sun.

16:00(EDT) On the strategy for supervised cell type identification in single-cell RNA-seq

Wenjing Ma, Kenong Su and ♦*Hao Wu*. Emory University

16:25(EDT) Semi-supervised learning for Single Cell Multi-Omics Data

Wei Chen. University of Pittsburgh

16:50(EDT) A graph neural network model to estimate cell-wise metabolic flux using single cell RNA-seq data

Wennan Chang[1]*, Norah Alghamdi*[2]*, Sha Cao*[2] *and* ♦*Chi Zhang*[2]. [1]Purdue University [2]Indiana University

17:15(EDT) Scaffold: a data generation simulation framework for single-cell RNA-seq data

Rhonda Bacher. University of Florida

17:40(EDT) Floor Discussion.

Sep. 15 10:20-noon (EDT)

Session 96: Semiparametric statistical inference and application

Organizer: Jing Yang.
Chair: Jing Yang.

10:20(EDT) The profile likelihood based statistical inference
Ziqi Chen. East China Normal University

10:45(EDT) Classified Generalized Linear Mixed Model Prediction Incorporating Pseudo-prior Information
♦*Haiqiang Ma*[1] *and Jiming Jiang*[2]. [1]Jiangxi University of Finance and Economics [2]University of California, Davis

11:10(EDT) Nonparametric Regression with Covariates Subject to Dependent Censoring
Hui Jiang[1], ♦*Lei Huang*[2] *and Yingcun Xia*[3]. [1]Huazhong University of Science and Technology [2]Southwest Jiaotong University [3]National University of Singapore

11:35(EDT) Floor Discussion.

Session 97: Advances in High-dimensional Statistics

Organizer: Anru Zhang.
Chair: Anru Zhang.

10:20(EDT) Variable Selection for Frechet Regression
Yichao Wu. University of Illinois at Chicago

10:45(EDT) Understanding Generalization in Deep Learning via Tensor Methods
Jingling Li[1], *Yanchao Sun*[1], *Jiahao Su*[1], *Taiji Suzuki*[2] *and* ♦*Furong Huang*[1]. [1]UMD [2]Riken, Japan

11:10(EDT) Bidimensional Linked Matrix Decomposition for Pan-Omics Pan-Cancer Analysis
ERIC LOCK. UNIVERSITY OF MINNESOTA

11:35(EDT) Tensor Factor Model Estimation by Iterative Projection
Yuefeng Han[1], *Rong Chen*[1], ♦*Dan Yang*[2] *and Cun-Hui Zhang*[1]. [1]Rutgers University [2]The University of Hong Kong

12:00(EDT) Floor Discussion.

Session 98: Factor Analysis and Random Matrix Theory

Organizer: Tracy Ke.
Chair: Kaizheng Wang.

10:20(EDT) Eigen selection in spectral clustering: a theory guided practice
Xiao Han[1], *Xin Tong*[2] *and* ♦*Yingying Fan*[2]. [1]USTC [2]USC

10:45(EDT) Selecting the number of components in PCA via random signflips
David Hong, Yue Sheng and ♦*Edgar Dobriban*. UPenn

11:10(EDT) An Lp theory of PCA and spectral clustering
Emmanuel Abbe[1], *Jianqing Fan*[2] *and* ♦*Kaizheng Wang*[3]. [1]EPFL [2]Princeton University [3]Columbia University

11:35(EDT) Estimation of spectra of high-dimensional separable covariance matrices
♦*Lili Wang*[1] *and Debashis Paul*[2]. [1]Zhejiang Gongshang University [2]University of California, Davis

12:00(EDT) Floor Discussion.

Session 99: Analyses of Electronic Health Records

Organizer: Yun Li.
Chair: Yun Li.

10:20(EDT) Statistical inference for natural language processing algorithms when predicting type 2 diabetes using electronic health record notes
Brian Egleston[1] *and* ♦*Ashis Chanda*[2]. [1]Fox Chase Cancer Center [2]Temple University

10:45(EDT) Prediction of relapse in chronic disease using fragmented medical records
Jiasheng Shi and ♦*Jing Huang*. University of Pennsylvania

11:10(EDT) Handling irregular observation in longitudinal studies using electronic health records
Eleanor Pullenayegum. The Hospital for Sick Children

11:35(EDT) Statistical Methods for Phenotyping with Positive-Only Electronic Health Record Data
Jinbo Chen. University of Pennsylvania

12:00(EDT) Floor Discussion.

Session 100: Integrative multi-omics inference

Organizer: Michael Love.
Chair: Michael Love.

10:20(EDT) Double-matched matrix decomposition for multi-view data
Dongbang Yuan and ♦*Irina Gaynanova*. Texas A&M University

10:45(EDT) Deep IDA: A Deep Learning Method for Integrative Discriminant Analysis of Multi-View Data with Feature Ranking
Jiuzhou Wang and ♦*Sandra Safo*. University of Minnesota

11:10(EDT) Integrating multi-omics data for causal inference
♦*Dan Zhou*[1] *and Eric Gamazon*[2]. [1]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA [2]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA; Clare Hall, University of Cambridge, Cambridge, UK; MRC Epidemiology Unit, University of Cambridge, Cambridge, UK

11:35(EDT) Floor Discussion.

Session 101: New advances of statistical inference for high-dimensional data

Organizer: Yuan Ke.
Chair: Yuan Ke.

10:20(EDT) Power-enhanced simultaneous test of high-dimensional mean vectors and covariance matrices with application to gene-set testing
♦*Xiufan Yu*[1], *Danning Li*[2], *Lingzhou Xue*[3] *and Runze Li*[3]. [1]University of Notre Dame [2]Northeast Normal University [3]Pennsylvania State University

10:45(EDT) A Distribution-Free Independence Test for High Dimension Data
♦*Zhanrui Cai, Jing Lei and Kathryn Roeder*. Carnegie Mellon University

11:10(EDT) Optimal sampling designs for online estimation of streaming multi-dimensional time series
Shuyang Bai[1], ♦*Rui Xie*[2] *and Ping Ma*[1]. [1]University of Georgia [2]University of Central Florida

11:35(EDT) Multiple autocovariance change-points detection for high-dimensional time series
Di Xiao[1], ♦*Haotian Xu*[2], *Yuan Ke*[1] *and Jeongyoun Ahn*[1]. [1]University of Georgia [2]University of Warwick

12:00(EDT) Floor Discussion.

Session 102: Recent Advances in Analysis of Data with Measurement Errors
Organizer: Yu-Jen Cheng.
Chair: Yu-Jen Cheng.

10:20(EDT) Methods for diagnostic accuracy with biomarker measurement error
♦*Ching-Yun Wang and Ziding Feng*. Fred Hutchinson Cancer Research Center

10:45(EDT) Robust estimation of the causal risk difference with misclassified outcome data
♦*Di Shu*[1] *and Grace Yi*[2]. [1]University of Pennsylvania & Children's Hospital of Philadelphia [2]University of Western Ontario

11:10(EDT) Addressing measurement error in random forests using quantitative bias analysis
Tammy Jiang. Boston University School of Public Health

11:35(EDT) Corrected Score Methods for Recurrent Event Data with Covariate Measurement Error
Hsiang Yu[1], ♦*Yu-Jen Cheng*[2] *and Ching-Yun Wang*[3]. [1]Taiwan Semiconductor Manufacturing [2]National Tsing Hua University [3]Fred Hutchinson Cancer Research Center

12:00(EDT) Floor Discussion.

Session 103: Modern streaming Data Analysis: anomaly detection and applications
Organizer: Ruizhi Zhang.
Chair: Ruizhi Zhang.

10:20(EDT) Solar Radiation Anomaly Events Modeling Using Spatial-Temporal Mutually Interactive Processes
Minghe Zhang[1], *Chen Xu*[1], *Andy Sun*[1], *Feng Qiu*[2] *and* ♦*Yao Xie*[1]. [1]Georgia Institute of Technology [2]Argonne National Lab

10:45(EDT) A Change-Point Marginal Regression Model for Stock Returns' Tail Exceedance under Market Variations
Haipeng Xing. Stony Brook University

11:10(EDT) Does enforcing fairness mitigate biases caused by subpopulation shift?
Subha Maity[1], *Debarghya Mukherjee*[1], *Mikhail Yurochkin*[2] *and* ♦*Yuekai Sun*[1]. [1]University of Michigan [2]MIT-IBM Watson AI Lab

11:35(EDT) Homeostasis phenomenon in conformal prediction and predictive distribution functions
Minge Xie. Rutgers University

12:00(EDT) Floor Discussion.

Session 104: Enhancing Causal Inference Using Genetics Data: Mendelian Randomization
Organizer: Ting Ye.
Chair: Ting Ye.

10:20(EDT) Exploiting public GWAS databases to identify and adjust for heritable confounders in Mendelian randomization studies.
Jean Morrison. University of Michigan

10:45(EDT) Mendelian Randomization Analysis Using Multiple Biomarkers of an Underlying Common Exposure
♦*Jin Jin*[1], *Guanghao Qi*[2], *Zhi Yu*[3] *and Nilanjan Chatterjee*[1]. [1]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University [2]Department of Biomedical Engineering, Johns Hopkins University [3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard

11:10(EDT) Almost exact Mendelian randomization
♦*Matthew Tudball*[1] *and Qingyuan Zhao*[2]. [1]University of Bristol [2]University of Cambridge

11:35(EDT) Floor Discussion.

Session 105: Recent Development in Functional Data Analysis
Organizer: Zhenhua Lin.
Chair: Zhenhua Lin.

10:20(EDT) In-game win probabilities for the National Rugby League
♦*Tianyu Guan*[1], *Robert Nguyen*[2], *Jiguo Cao*[3] *and Tim Swartz*[3]. [1]Brock University [2]University of New South Wales [3]Simon Fraser University

10:45(EDT) Temporal-dependent Principal Component Analysis of Two-dimensional Functional Data
♦*Kejun He*[1], *Shirun Shen*[1] *and Lan Zhou*[2]. [1]Renmin University of China [2]Texas A&M University

11:10(EDT) Inference on Linear Models for Functional Data via Bootstrapping Max Statistics
Yinan Lin. National Univerisity of Singapore

11:35(EDT) Basis Expansions for Functional Snippets
Zhenhua Lin[1], *Jane-Ling Wang*[2] *and* ♦*Qixian Zhong*[3]. [1]National University of Singapor [2]University of California [3]Xiamen University

12:00(EDT) Floor Discussion.

Session 106: New Advances in High-Dimensional Time Series Analysis
Organizer: Yao Zheng.
Chair: Yao Zheng.

10:20(EDT) Learning Financial Network with Focally Sparse Structure
♦*Weining Wang*[1], *Chen Huang*[2] *and Victor Chernozhukov*[3]. [1]U of York [2]U of aarhus [3]MIT

10:45(EDT) Online Inference with Stochastic Gradient Descent via Random Scaling
Simon Lee[1], *Matt Seo*[2], *Youngki Shin*[3] *and* ♦*Yuan Liao*[4]. [1]Columbia [2]Seoul national university [3]McMaster University [4]Rutgers

11:10(EDT) Title: Forecasting time series with Wasserstein GANs
Moritz Haas and ♦*Stefan Richter*. Heidelberg University

11:35(EDT) Floor Discussion.

Session 107: Recent Developments for network data analysis: Theory, Method, and Application
Organizer: Guanyu Hu.
Chair: Guanyu Hu.

10:20(EDT) Euclidean Representation of Low-Rank Matrices and Its Statistical Applications
Fangzheng Xie. Indiana University

10:45(EDT) Block Mean-mean-field variational inference for dynamic latent space models
♦*Peng Zhao, Debdeep Pati, Anirban Bhattacharya and Bani Mallick*. Texas A&M University

11:10(EDT) Network Functional Varying Coefficient Model
Xuening Zhu. Fudan University

11:35(EDT) Finite Mixtures of ERGMs for Modeling Ensembles of Networks
♦*Fan Yin*[1]*, Weining Shen*[2] *and Carter Butts*[2]. [1]Microsoft [2]University of California, Irvine

12:00(EDT) Floor Discussion.

Session 108: Recent advances in Bayesian methodology for complex data
Organizer: Quan Zhou.
Chair: Quan Zhou.

10:20(EDT) Data integration using hierarchical Gaussian process models under shape constraints
♦*Shuang Zhou*[1]*, Debdeep Pati*[2] *and Anirban Bhattacharya*[2]. [1]Arizona State University [2]Texas A&M University

10:45(EDT) Bayesian modeling of spatial molecular profiling data
Qiwei Li. The University of Texas at Dallas

11:10(EDT) A large-scale Bayesian spatial extremes modeling approach with application to wind extremes in Saudi Arabia
♦*Wanfang Chen*[1]*, Stefano Castruccio*[2] *and Marc Genton*[3]. [1]East China Normal University [2]University of Notre Dame [3]King Abdullah University of Science and Technology

11:35(EDT) Dimension-free mixing for high-dimensional Bayesian variable selection
♦*Quan Zhou*[1]*, Jun Yang*[2]*, Dootika Vats*[3]*, Gareth Roberts*[4] *and Jeffrey Rosenthal*[5]. [1]Texas A&M University [2]Oxford University [3]Indian Institute of Technology Kanpur [4]University of Warwick [5]University of Toronto

12:00(EDT) Floor Discussion.

Session 109: Recent Advances in Linear Mixed Models and Applications
Organizer: Min Wang.
Chair: Min Wang.

10:20(EDT) Informative g-priors for Mixed Models
Yu-Fang Chien[1]*,* ♦*Haiming Zhou*[1]*, Timothy Hanson*[2] *and Ted Lystig*[2]. [1]Northern Illinois University [2]Medtronic

10:45(EDT) Bayesian quantile semiparametric mixed-effects double regression models
♦*Min Wang*[1]*, Duo Zhang*[2]*, Liucang Wu*[3] *and Keying Ye*[1]. [1]University of Texas at San Antonio [2]Michigan Tech University [3]Kunming University of Science and Technology

11:10(EDT) Small area estimation with subgroup analysis
♦*Xin Wang*[1] *and Zhengyuan Zhu*[2]. [1]Miami University [2]Iowa State University

11:35(EDT) Selecting Mixed Effects Models using Penalized Profile REML with Application to the Cohort Study of HIV
Juming Pan. 42 WATSON DR

12:00(EDT) Floor Discussion.

Abstracts

Session 1: Student Paper Competition Winners

Query-augmented Active Metric Learning

♦*Yujia Deng*[1]*, Yubai Yuan*[2]*, Haoda Fu*[3] *and Annie Qu*[2]
[1]University of Illinois, Urbana-Champaign
[2]University of California, Irvine
[3]Eli Lilly and Company
yujiad2@illinois.edu

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

Transportation of Area Under the ROC curve to a Target Population

♦*Bing Li*[1]*, Issa Dahabreh*[2]*, Constantine Gatsonis*[1] *and Jon Steingrimsson*[1]
[1]Brown University
[2]Harvard University
bing_li1@brown.edu

We develop methods for estimating the area under the ROC curve (AUC) of a prediction model in a target population that differs from the source population population used for original model development. We focus on the setting where outcome and covariate data are available from the source population, but only covariate data are available from the target population. If covariates that affect model AUC are differently distributed between the source and target population, AUC estimators that only use data from the source population are biased for the target population AUC. We provide identifiability conditions and results under which the target population AUC is identifiable. We develop three estimators for the target population AUC and show that they are consistent and asymptotically normal. We evaluate their finite-sample performance using simulations and we apply the to estimate the AUC of a lung cancer risk prediction model using source population data from the National Lung Screening Trial (NLST) and target population data from the National Health and Nutrition Examination Survey (NHANES).

Scalable community detection in massive networks via predictive inference

♦*Subhankar Bhadra*[1]*, Marianna Pensky*[2] *and Srijan Sengupta*[1]
[1]NC State University
[2]University of Central Florida
sbhadra@ncsu.edu

Identification of community structure in empirical networks has been of particular interest in the statistics literature. Community detection for massive networks with millions of nodes has always been a topic of great relevance, since most of the existing standard community detection algorithms take long time to run because of the large matrix computations involved. In this paper, we propose a novel community detection algorithm using predictive inference, where we use any statistically sound community detection algorithm to classify a sub-graph of the data and extend the classification to the rest of the nodes borrowing information from the assumed model. Our simulation studies show that using our algorithm makes huge improvement in terms of run-time losing very negligible amount of accuracy.

The Promises of Parallel Outcomes

♦*Ying Zhou, Dehan Kong and Linbo Wang*
University of Toronto
yingx.zhou@mail.utoronto.ca

Unobserved confounding presents a major threat to the validity of causal inference from observational studies. In this paper, we introduce a novel framework that leverages the information in multiple parallel outcomes for identification and estimation of causal effects. Under a conditional independence structure among multiple parallel outcomes, we achieve nonparametric identification with at least three parallel outcomes. We further show that under a set of linear structural equation models, causal inference is possible with two parallel outcomes. We develop accompanying estimating procedures and evaluate their finite sample performance through simulation studies and a data application studying the causal effect of the tau protein level on various types of behavioral deficits.

A Wavelet-Based Independence Test for Functional Data with an Application to MEG Functional Connectivity

♦*Rui Miao*[1]*, Xiaoke Zhang*[1] *and Raymond Wong*[2]
[1]The George Washington University
[2]Texas A&M University
ruimiao@gwu.edu

Measuring and testing the dependency between multiple random functions is of- ten an important task in functional data analysis. In the literature, a model-based method relies on a model which is subject to the risk of model misspecification, while a model-free method only provides a correlation measure which is inadequate to test independence. In this paper, we adopt the Hilbert-Schmidt Independence Criterion (HSIC) to measure the dependency between two random functions. We develop a two- step procedure by first pre-smoothing each function based on its discrete and noisy measurements and then applying the HSIC to recovered functions. To ensure the compatibility between the two steps such that the effect of the pre-smoothing error on the subsequent HSIC is asymptotically negligible when the data are densely measured, we propose a new wavelet thresholding method for pre-smoothing and to use Besov-norm-induced kernels for HSIC. We also provide the corresponding asymptotic analysis. The superior numerical performance of the proposed method over existing ones is demonstrated in a simulation study. Moreover, in a magnetoencephalography (MEG) data

application, the functional connectivity patterns identified by the proposed method are more anatomically interpretable than those by existing methods.

Linear Models with Distributed Data and Communication-Efficient Estimator: Limiting Distribution and Its Approximation

Ziyan Yin
Temple University
tuj64107@temple.edu

In current large-scale practical investigations, it is common that data are distributed in multiple parallel centers. For reasons including confidentiality concerns, the data from every center are not all available when conducting statistical analysis. To solve the data-enabled problems more efficiently in this scenario, a class of communication-efficient methods are being actively developed. We investigate in this study the limiting distribution of the estimators that requires a one-step updating, combining data information transferred from all parallel centers. Our analysis reveals that the number of centers is a critical issue. In particular, when the number of centers is not small compared with the sample size at the local center, a non-negligible impact occurs in the limiting distribution. Moreover, we find that the limiting distribution itself is interesting – it is a mixture involving a product of normal random variables. As a result, conventional methods for statistical inferences by only incorporating the variance estimations may be severely biased – leading to serious under-coverage. Based on our analysis, we then propose a multiplier-bootstrap method for approximating the distribution of the estimator. To meet the challenges in high-dimensional problems, we further develop a data-splitting and re-fitting strategy to handle the limiting distribution therein.

Session 2: Recent Advances in Deep Learning

Advances of Momentum in Optimization Algorithm and Neural Architecture Design

Bao Wang
University of Utah
bwang@sci.utah.edu

We will present a few recent results on leveraging momentum techniques to improve stochastic optimization and neural architecture design. First, designing deep neural networks is an art that often involves an expensive search over candidate architectures. To overcome this for recurrent neural nets (RNNs), we establish a connection between the hidden state dynamics in an RNN and gradient descent (GD). We then integrate momentum into this framework and propose a new family of RNNs, called *MomentumRNNs*. We theoretically prove and numerically demonstrate that MomentumRNNs alleviate the vanishing gradient issue in training RNNs. Also, we show the empirical advantage of the momentum enhanced RNNs over the baseline models. Second, we will present the recent advances of adaptive momentum in accelerating the stochastic gradient descent (SGD). The adaptive momentum assisted SGD remarkably improves the deep neural network training in terms of acceleration and improved generalization and significantly reduces the effort for hyperparameter tuning.

Diff-ResNet: Diffusion Augmented Neural Network

Tangjun Wang, Chenglong Bao and ♦*Zuoqiang Shi*
Tsinghua University
zqshi@tsinghua.edu.cn

Interpreting deep neural networks from the ordinary differential equations (ODEs) perspective has inspired many efficient and robust network architectures. However, existing ODE based approaches ignore the relationship among data points, which is a critical component in many problems including few-shot learning and semi-supervised learning. In this talk,we propose a novel diffusion residual network (Diff-ResNet) to strengthen the interactions among data points. Under the structured data assumption, it is proved that the diffusion mechanism can decrease the distance-diameter ratio that improves the separability of inter-class points and reduces the distance among local intra-class points. This property can be easily adopted by the residual networks for constructing the separable hyperplanes. The synthetic binary classification experiments demonstrate the effectiveness of the proposed diffusion mechanism. Moreover, extensive experiments of few-shot image classification and semi-supervised graph node classification in various datasets validate the advantages of the proposed Diff-ResNet over existing few-shot learning methods.

SODEN: A Scalable Continuous-Time Survival Model through Ordinary Differential Equation Networks

♦*Weijing Tang, Jiaqi Ma, Qiaozhu Mei and Ji Zhu*
University of Michigan
WEIJTANG@UMICH.EDU

In this paper, we propose a flexible model for survival analysis using neural networks along with scalable optimization algorithms. One key technical challenge for directly applying maximum likelihood estimation (MLE) to censored data is that evaluating the objective function and its gradients with respect to model parameters requires the calculation of integrals. To address this challenge, we recognize from a novel perspective that the MLE for censored data can be viewed as a differential-equation constrained optimization problem. Following this connection, we model the distribution of event time through an ordinary differential equation and utilize efficient ODE solvers and adjoint sensitivity analysis to numerically evaluate the likelihood and the gradients. Using this approach, we are able to 1) provide a broad family of continuous-time survival distributions without strong structural assumptions, 2) obtain powerful feature representations using neural networks, and 3) allow efficient estimation of the model in large-scale applications using stochastic gradient descent. Through both simulation studies and real-world data examples, we demonstrate the effectiveness of the proposed method in comparison to existing state-of-the-art deep learning survival analysis models.

A graph deep learning model based on neural ordinary differential equations

♦*Yuanhao Liu, Changpeng Lu and Sijian Wang*
Rutgers University
yl1398@scarletmail.rutgers.edu

Protein classification is an important problem in biology. Proteins can be viewed as graphs, whose nodes are amino acids with some node features and whose edges are interaction energy between nodes. Thus, protein classification problem can be viewed as graph classification problem. In literature, there are two main schemes of approaches for solving this problem. One is pure data driven deep learning approach such as graph convolutional network (GCN). The other one is based on theories of physics and chemistry, using ordinary equation to derive the properties of proteins. In this work, we combine these two kinds of approaches and propose connection diffusion model, which is a family of deep learning models based on diffusively coupled nonlinear oscillator system. We also

show that, by specifying functions in our model, GCN and ResNet GCN are included as two special cases of this family. We also apply our model on some benchmark datasets and make comparison with GCN model.

Session 3: Missing Data Method Development and Applications

Semiparametrically efficient approach for regression analysis with missing response and a hypothesis testing procedure for the missing data mechanism

♦*Qinglong Tian*[1]*, Jiwei Zhao*[1] *and Donglin Zeng*[2]
[1]University of Wisconsin - Madison
[2]The University of North Carolina at Chapel Hill
qtian27@wisc.edu

This paper considers a parametric regression model with missing responses. The missing data mechanism is left unspecified and could potentially be missing not at random. We first study the identifiability issue in this model and propose a general condition such that all the unknown components are identifiable. Then we propose a semiparametrically efficient estimation approach for the regression analysis, develop an EM-based algorithm for implementation, and establish the asymptotic theory of the proposed method. Based on our proposal, we further develop a hypothesis testing procedure for testing whether the missing data mechanism is indeed missing not at random. We conduct comprehensive simulation studies and also apply the proposed method to a real data to demonstrate the usefulness of the proposed method. This is based on a joint work with Donglin Zeng and Jiwei Zhao.

A practical solution for missing data in clinical outcome-a simulation study of multiple imputations in a real-world patient registry

♦*Kim Hung Lo*[1]*, Yiting Wang*[2]*, Chunyan Liu*[3] *and Nanhua Zhang*[3]
[1]Clinical Biostatistics, Janssen R&D
[2]Global Epidemiology, Janssen R&D
[3]Division of Biostatistics & Epidemiology, Cincinnati Children's Hospital
KLo2@its.jnj.com

Patient registries are increasingly used to generate real-world evidence to assess product effectiveness for regulatory decision-making. Missing/incomplete data of clinical outcome measure often present a major hurdle. A simulation study was conducted to evaluate multiple imputations (MI) as a practical solution for handling missing data in clinical remission status at week-52 after the initiation of a biologic agent in pediatric patients with Crohn's disease (CD) from ImproveCareNow (ICN), the world's largest registry of pediatric patients with inflammatory bowel disease (IBD). Our strategies integrated both statistical and clinical perspectives, including the choices of continuous versus categorized candidate predictors; predictor selection models of individual components and the summary outcome measure, as well as missingness of the outcome; and visit window consideration. Starting from a complete dataset with known outcome measures and thus known "true" remission rate, different levels of missingness were imposed and MI was performed. The performance of the MI method was evaluated by relative bias, defined as (estimated - true)/true, and coverage, defined as the proportion of covering the "true" remission rate by 95% confidence intervals in the replications. Results showed excellent performance of the MI method, which essentially removed all bias compared with "complete-case" analysis, and coverage with the MI method was 100% across all simulated scenarios.

Model-based multiple imputation method for multilevel regression models with left-censored data and derived predictors

♦*Peixin Xu*[1]*, Aimin Chen*[2]*, Nanhua Zhang*[3] *and Changchun Xie*[1]
[1]University of Cincinnati College of Medicine
[2]University of Pennsylvania Perelman School of Medicine
[3]Cincinnati Children's Hospital Medical Center & University of Cincinnati College of Medicine
xupi@mail.uc.edu

Multilevel regression with predictors that are left-censored due to the limits of detection (LOD) are common in biomedical and epidemiological studies. One of the widely used methods replaces missing data below LOD with a single and fixed value such as LOD, LOD/2 or LOD/v2; this method ignores the variation of missing data and thus results in potentially biased parameter estimates with underestimated variances. On the other hand, traditional multiple imputation (MI) methods such as joint modeling and fully conditional specification do not guarantee that imputed values will fall below the LOD. When the multilevel regression model contains derived predictors such as interaction or quadratic terms, chained-equations based MI methods result in biased parameter estimates since the "reverse regression imputation" ignores the relationship between outcome and derived terms. We propose a Bayesian model-based multiple imputation method based on Gibbs sampling that can accommodate left-censored data below LOD as well as derived predictors. Missing data below LOD will be imputed by draws from a truncated normal distribution that is proportional to the product of two terms: (i) the analysis model and (ii) a conditional model for the target predictor given the rest of covariates. The proposed method can also handle partially observed outcome assuming missing at random. Our simulation studies demonstrate that our method outperforms traditional imputation methods for left-censored data in parameter estimation. The proposed method is then applied to a study of organophosphate ester (OPE) flame retardants to assess the association with health outcomes.

Session 4: Recent advances in methodologies for omics data analysis

NanoSplicer: Accurate identification of splice junctions using Oxford Nanopore sequencing

Yupei You[1]*, Michael Clark*[2] *and* ♦*Heejung Shim*[1]
[1]School of Mathematics and Statistics and Melbourne Integrative Genomics, The University of Melbourne, Parkville, Australia
[2]Centre for Stem Cell Systems, Department of Anatomy and Neuroscience, The University of Melbourne, Parkville, Australia
heejung.shim@unimelb.edu.au

Alternative splicing is an essential mechanism that enables a single gene to produce multiple mRNA products (called isoforms). Oxford nanopore sequencing produces long reads that have natural advantages for characterising isoforms. Alternative splicing can not only add or skip entire exons, but can also vary exon boundaries by selecting different nucleotides as splice sites. However, accurately identifying the latter is challenging for nanopore reads, as exon boundaries are often unclear due to the high error rate. One existing solution is polishing nanopore reads with short reads. While feasible, this approach requires both short and long reads, which adds considerable expense. Furthermore, isoform-distinguishing short

reads are not always available (e.g. in 10X scRNAseq). Therefore, a method that could accurately identify splice junctions solely from nanopore data would have numerous advantages. We developed a method called "NanoSplicer" to identify splice junctions using only nanopore sequencing data. Nanopore sequencing records changes in electrical current when a DNA or RNA strand is traversing through a pore. This raw signal (known as a squiggle) is then basecalled by computational methods, but this process is error-prone. Instead of looking at the basecalled reads only, we also use the squiggle to identify splice junctions. Using both synthetic mRNAs with known splice junctions and biological real data, we show that our method improves upon reads-only methods, especially for reads with high basecalling errors.

Cross-Platform Omics Prediction procedure: a statistical framework for implementing precision medicine

Jean Yang
The University of Sydney,
jean.yang@sydney.edu.au

In the modern era of precision medicine, molecular signatures identified from biotechnological advancement hold great promise to better guide clinical decisions and improve patient outcomes. However, current approaches in biomarker identification and risk model construction are often platform and site-specific which poses a significant challenge to wide use in clinical practice. Here, we present a novel Cross-Platform Omics Prediction (CPOP) procedure that identifies biomarkers and transferable models that are platform-independent and stable across time and studies. We demonstrate its utility by building a risk model for early-stage metastatic melanoma (stage III) and validate the procedure in two publicly available datasets and on an independent cohort (n = 46) of retrospective samples. We also demonstrate the generalisability of this algorithm by applying CPOP to two additional diseases with different experimental designs. The ability of CPOP to be used prospectively and across multiple sites is a significant step forward to facilitating the wide-spread use of omics signatures as precise biomarkers in clinical practice.

RUV-III-NB: Removing Unwanted Variation from High-throughput Single-Cell RNA-seq Data

♦*Agus Salim*[1]*, Ramyar Molania*[2]*, Jianan Wang*[2] *and Terence Speed*[2]
[1]University of Melbourne
[2]Walter Eliza Hall Institute of Medical Research
salim.a@unimelb.edu.au

Due to advances in technology, high-throughput RNA measurement at single-cell level is now available. More and more studies are using this technology, which has already led to identification of novel cell types and guiding immunotherapy for cancer treatment. However, this data is highly-complex with a considerable amount of unwanted technical variation that can obscure our ability to detect important biological signals. Current methods for removing these unwanted variations exist but they are not always effective as they tend to remove biological signals whilst trying to remove the unwanted variation. We develop a statistical method for removing unwanted variation from single-cell RNA sequencing data. The statistical method models the raw sequencing count using zero-inflated negative binomial (ZINB) distribution. Using technical replicates and control features (genes), the method estimates the unwanted factors via a modified iterative reweighted least squares (IRLS) algorithm and remove the effect of these unwanted factors from the raw sequencing count using randomized quantile residuals approach. Using simulated and real datasets, we compare the method to leading methods for removing unwanted variation in single-cell RNA sequencing data and demonstrate the comparative advantage of our method in removing unwanted factors while retaining important biological signals. The method is implemented as an R package and available from the following Github site: https://github.com/limfuxing/ruvIIInb

Multiple Testing of One-sided Hypotheses under General Dependence

Seonghun Cho[1]*,* ♦*Youngrae Kim*[1]*, Hyungwon Choi*[2]*, Johan Lim*[1]*, DoHwan Park*[3] *and Woncheol Jang*[1]
[1]Department of Statistics, Seoul National University
[2]School of Medicine, National University of Singapore
[3]Department of Mathematics and Statistics, University of Maryland
acarlos93@snu.ac.kr

In this work, we propose a procedure, named DAB-PFA, to test many one-sided hypotheses simultaneously under the general dependency of test statistics. The one-sided hypothesis in the multiple testing makes the empirical null distribution (or p-values) be conservative and further introduces a significant loss in power if we do not take this account appropriately. In this work, we use the principal factor approximation by Fan and Han (2017) to account for the dependency among test statistics, and propose to adaptively discard statistics with small or large p-values in estimating the false discovery proportion (FDP) to resolve the conservativeness from the one-sided hypothesis. We theoretically prove the convergence of the estimated FDP by DAB- PFA to the true FDP and compute its rate. We also numerically compare the FDP control and power of the proposed DAB-PFA to existing procedures, Benjamini and Hochberg (1995), Efron (2004) and Wang and Fan (2017). Finally, we apply the proposed method to protein phosphorylation analysis of ovarian serous adenocarcinoma to identify protein modification levels uniquely elevated in each of the five molecular subtypes.

Session 5: Statistical methods for microbiome data analysis

mbImpute: an accurate and robust imputation method for microbiome data

♦*Ruochen Jiang*[1]*, Wei Vivian Li*[2] *and Jessica Jingyi Li*[1]
[1]University of California, Los Angeles
[2]Rutgers school of Public Health
ruochenj@gmail.com

Microbiome studies have gained increased attention since many discoveries revealed connections between human microbiome compositions and diseases. A critical challenge in microbiome research is that excess non-biological zeros distort taxon abundances, complicate data analysis, and jeopardize the reliability of scientific discoveries. To address this issue, we propose the first imputation method, mbImpute, to identify and recover likely non-biological zeros by borrowing information jointly from similar samples, similar taxa, and optional metadata including sample covariates and taxon phylogeny. Comprehensive simulations verified that mbImpute achieved better imputation accuracy under multiple measures than five state-of-the-art imputation methods designed for non-microbiome data. In real data applications, we demonstrate that mbImpute improved the power and reproducibility of identifying disease-related taxa from microbiome data of type 2 diabetes and colorectal cancer.

Linear Models for Differential Abundance Analysis of Microbiome Compositional Data

♦*Jun Chen*[1] *and Xianyang Zhang*[2]
[1]Mayo Clinic
[2]Texas 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.

Variable selection analysis in small-sample microbiome compositional data

Arun Srinivasan, Lingzhou Xue and ♦*Xiang Zhan*
Penn State University
xiangzhan9@gmail.com

A critical task in microbiome studies is to identify microbial features that are associated with outcomes. Classic statistical variable selection methods such as Lasso, SCAD and MCP are all versatile and enjoying nice asymptotic properties. Yet, they are not quite applicable to microbiome data partially due to compositionality and small sample size. Motivated by fine-mapping of the microbiome, we propose a two-step compositional knockoff filter (CKF) to provide the effective finite-sample false discovery rate (FDR) control in high-dimensional linear log-contrast regression analysis of microbiome compositional data. In the first step, we employ the compositional screening procedure to remove insignificant microbial taxa while retaining the essential sum-to-zero constraint. In the second step, we extend the knockoff filter to identify the significant microbial taxa in the sparse regression model for compositional data. We study the asymptotic properties of the proposed two-step procedure, including both sure screening and effective false discovery control. The potential usefulness of the proposed method is also illustrated with both simulation studies and real data applications.

Microbial Trend Analysis in Longitudinal Microbiome Study

Chan Wang[1]*, Jiyuan Hu*[1]*, Martin Blaser*[2] *and* ♦*Huilin Li*[1]
[1]New York University
[2]Rutgers University
Huilin.Li@nyulangone.org

The human microbiome is inherently dynamic and its dynamic nature plays a critical role in maintaining health and driving disease. With an increasing number of longitudinal microbiome studies, scientists are eager to learn the comprehensive characterization of microbial dynamics and their implications to the health and disease-related phenotypes. However, due to the challenging structure of longitudinal microbiome data, few analytic methods are available to characterize the microbial dynamics over time. We propose a microbial trend analysis (MTA) framework for the high-dimensional and phylogenetically-based longitudinal microbiome data. In particular, MTA can perform three tasks: 1) capture the common microbial dynamic trends for a group of subjects at the community level and identify the dominant taxa; 2) examine whether or not the microbial overall dynamic trends are significantly different between groups; 3) classify an individual subject based on its longitudinal microbial profiling. Our extensive simulations demonstrate that the proposed MTA framework is robust and powerful in hypothesis testing, taxon identification, and subject classification. Our real data analyses further illustrate the utility of MTA through a longitudinal study in mice. In conclusions, the proposed MTA framework is an attractive and effective tool in investigating dynamic microbial pattern from longitudinal microbiome studies.

Session 6: Modern Statistical Methods for Analyzing Time Series Data

Biclustering Approaches for High-Frequency Time Series

Haitao Liu[1]*, Jian Zou*[1] *and* ♦*Nalini Ravishanker*[2]
[1]Worcester Polytechnic Institute
[2]University of Connecticut
nalini.ravishanker@uconn.edu

Clustering a large number of time series into relatively homogeneous groups is a well-studied unsupervised learning technique that has been widely used for grouping financial instruments (say, stocks) based on their stochastic properties across the entire time period under consideration. However, clustering algorithms ignore the notion of co-clustering, i.e., grouping of stocks only within a subset of times rather than over the entire time period. Biclustering techniques are useful for simultaneously clustering rows and columns of a data matrix. Over the past two decades, there has been a proliferation of biclustering approaches and interest in this area continues to grow. It is useful to apply biclustering algorithms to a large set of long time series that occur in many application domains, such as the bio-sciences, finance, etc. This talk will give an overview on biclustering approaches, followed by descriptions of two algorithms that we have developed to bicluster intra-day stock returns time series over multiple trading days. The algorithms employ the mean residue score and mutual information as metrics. Through some post-biclustering analyses, we show how data analysts may make use of the biclustering results to study co-movement patterns in sets of stock returns within time blocks.

Bayesian Estimation of Time-varying models

Sayar Karmakar
University of Florida
sayarkarmakar@ufl.edu

A major motivation of time-varying models emanate from the field of econometrics but these are also prevalent in several other areas such as medical sciences, climatology etc. Whenever a time-series dataset is observed over a large period of time, it is natural to assume the coefficient parameters also vary over time. In this talk

we review such models for two specific applications: Modelling Poisson count data through time-varying auto egression and Modeling ARCH-GARCH through time-varying models. For the first, time-varying model is not very well-researched but the recent covid count data necessitates this problem to be well-studied. For the latter, although a lot has been done in the frequentist regime, it has received little attention in terms of developing Bayesian theory. Towards that, we propose a Bayesian approach to the estimation of such models and develop a computationally efficient MCMC algorithm based on Hamiltonian Monte Carlo (HMC) sampling. We also establish posterior contraction rates with increasing sample size in terms of the average Hellinger metric. The performance of our method is compared with frequentist estimates and estimates from the time constant analogs for both count data and ARCH-GARCH data. We conclude the talk by discussing two applications: a. the Covid-19 spread in NYC through the tvPoisson model and b. Predictive performance comparison of the Bayesian and frequentist tv(G)ARCHmodel applied on some real datasets.

Spectral methods for small sample time series: A complete periodogram approach

Sourav Das[1], ♦*Suhasini Subba Rao*[2] *and Junho Yang*[3]
[1]James Cook University
[2]Texas A&M University
[3]Texa A&M University
suhasini@tamu.edu

The periodogram is a widely used tool to analyze second order stationary time series. An attractive feature of the periodogram is that the expectation of the periodogram is approximately equal to the underlying spectral density of the time series. However, this is only an approximation, and it is well known that the periodogram has a finite sample bias, which can be severe in small samples. In this paper, we show that the bias arises because of the finite boundary of observation in one of the discrete Fourier transforms which is used in the construction of the periodogram. Moreover, we show that by using the best linear predictors of the time series over the boundary of observation we can obtain a "complete periodogram" that is an unbiased estimator of the spectral density. In practice, the "complete periodogram" cannot be evaluated as the best linear predictors are unknown. We propose a method for estimating the best linear predictors and prove that the resulting "estimated complete periodogram" has a smaller bias than the regular periodogram. The estimated complete periodogram and a tapered version of it are used to estimate parameters, which can be represented in terms of the integrated spectral density. We prove that the resulting estimators have a smaller bias than their regular periodogram counterparts. The proposed method is illustrated with simulations and real data.

Large Spectral Density Matrix Estimation by Adaptive Thresholding

Yiming Sun[1], *Yige Li*[2], *Amy Kuceyeski*[1] *and* ♦*Sumanta Basu*[1]
[1]Cornell University
[2]Harvard University
sumbose@cornell.edu

Spectral density matrix estimation of multivariate time series is a classical problem in time series and signal processing. In modern neuroscience, spectral density based metrics are commonly used for analyzing functional connectivity among brain regions. In this paper, we develop a non-asymptotic theory for regularized estimation of high-dimensional spectral density matrices of Gaussian and linear processes using adaptively thresholded versions of averaged periodograms. Our theoretical analysis ensures that consistent estimation of spectral density matrix of a p-dimensional time series using n samples is possible under high-dimensional regime $\log p = o(n)$ as long as the true spectral density is approximately sparse. A key technical component of our analysis is a new concentration inequality of average periodogram around its expectation, which is of independent interest. Our estimation consistency results complement existing results for shrinkage based estimators of multivariate spectral density, which require no assumption on sparsity but only ensure consistent estimation in a regime $p^2 = o(n)$. In addition, our proposed thresholding based estimators perform consistent and automatic edge selection when learning coherence networks among the components of a multivariate time series. We demonstrate the advantage of our estimators using simulation studies and a real data application on functional connectivity analysis with fMRI data.

Session 7: Modern machine learning: method and theory

On Function Approximation in Reinforcement Learning: Optimism in the Face of Large State Spaces

♦*Zhuoran Yang*[1], *Chi Jin*[1], *Zhaoran Wang*[2], *Mengdi Wang*[1] *and Michael Jordan*[3]
[1]Princeton
[2]Northwestern
[3]UC Berkeley
zy6@princeton.edu

The classical theory of reinforcement learning (RL) has focused on tabular and linear representations of value functions. Further progress hinges on combining RL with modern function approximators such as kernel functions and deep neural networks, and indeed there have been many empirical successes that have exploited such combinations in large-scale applications. There are profound challenges, however, in developing a theory to support this enterprise, most notably the need to take into consideration the exploration-exploitation tradeoff at the core of RL in conjunction with the computational and statistical tradeoffs that arise in modern function-approximation-based learning systems. We approach these challenges by studying an optimistic modification of the least-squares value iteration algorithm, in the context of the action-value function represented by a kernel function or an overparameterized neural network. We establish both polynomial runtime complexity and polynomial sample complexity for this algorithm, without additional assumptions on the data-generating model. In particular, we prove that the algorithm incurs a sublinear regret. Our regret bounds are independent of the number of states, a result which exhibits clearly the benefit of function approximation in RL.

The cost of privacy in generalized linear models: algorithms and optimal rate of convergence

Tony Cai[1], *Yichen Wang*[1] *and* ♦*Linjun Zhang*[2]
[1]University of Pennsylvania
[2]Rutgers University
linjun.zhang@rutgers.edu

Differentially private algorithms are proposed for parameter estimation in both low-dimensional and high-dimensional sparse generalized linear models (GLMs) by constructing private versions of projected gradient descent. We show that the proposed algorithms are nearly rate-optimal by characterizing their statistical performance and establishing privacy-constrained minimax lower bounds for GLMs. The lower bounds are obtained via a novel technique based on Stein's Lemma that generalizes the tracing attack technique for privacy-constrained lower bounds. This lower bound argument can be of independent interest as it applies to general para-

metric models. Simulated and real data experiments are conducted to demonstrate the numerical performance of our algorithms.

On the Statistical Properties of Adversarial Robust Estimators

♦*Yue Xing, Qifan Song and Guang Cheng*
Purdue University
xing49@purdue.edu

The development of machine/deep learning methods has led to breakthrough performances in modern AI applications. However, recent research reveals that these powerful models can be very vulnerable to undesriable perturbations in testing data input. This thus emphasizes the importance of studying adversarial robustness in deep learning. This presentation focuses on the statistical understanding of adversarial training. Specifically, we consider linear regression models and two-layer neural networks (with lazy training) using squared loss under low-dimensional and high-dimensional regimes. In the former regime, after overcoming the non-smoothness of adversarial training, the adversarial risk of the trained models can converge to the minimal adversarial risk, and the algorithmic stability also gets improved. These two improvements combined imply a better generalization performance. In the latter regime, we discover that data interpolation prevents the adversarially robust estimator from being consistent. Therefore, inspired by successes of LASSO, we incorporate the L1 penalty in the high dimensional adversarial learning and show that it leads to consistent adversarially robust estimation.

Session 8: Semiparametric and nonparametric methods in causal inference with multi- or high-dimensional confounders

Modeling the natural history of human diseases

♦*Yen-Tsung Huang and Ju-Sheng Hong*
Academia Sinica
ythuang@stat.sinica.edu.tw

Natural history of human diseases is comprised of several sequential milestones that are time-to-event by nature. For example, from hepatitis B infection to death, patients may experience intermediate events such as liver cirrhosis and liver cancer. The events of hepatitis, cirrhosis, cancer and death have a sequential order and are subject to right censoring; moreover, the latter events may mask the former ones. By casting the natural history of human diseases in the framework of causal mediation modeling, we set up a mediation model with intermediate events and a terminal event, respectively as the mediators and the outcome. We define the interventional analogue of path-specific effects (iPSEs) as the effect of an exposure on a terminal event mediated by any combination of intermediate events, including not through any of the events. We derive expression of a counterfactual hazard under sequential ignorability. We construct composite nonparametric likelihood and obtain a Nelson-Aalen type of estimators for the counterfactual hazard and iPSEs. We establish asymptotic unbiasedness, uniform consistency and weak convergence for the proposed estimators. Numerical studies including simulation and data application are presented to illustrate the finite sample performance and utility of the proposed method.

Inference for algorithm-agnostic variable importance

Brian Williamson[1]*, Peter Gilbert*[2]*, Noah Simon*[3] *and* ♦*Marco Carone*[3]
[1]Kaiser Permanente Washington Health Research Institute
[2]Fred Hutchinson Cancer Research Center
[3]University of Washington
mcarone@uw.edu

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, that is, 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, we propose a general framework for nonparametric inference on interpretable algorithm-agnostic variable importance. We define variable importance as a population-level contrast between the oracle predictiveness of all available features versus all features except those under consideration. We then propose a nonparametric efficient estimation procedure that allows the construction of valid confidence intervals and tests, even when machine learning techniques are used. We discuss the use of this general framework on a causal measure of variable importance.

Optimal tests of the composite null hypothesis arising in mediation analysis

♦*Caleb Miles*[1] *and Antoine Chambaz*[2]
[1]Columbia University
[2]Université de Paris
cm3825@cumc.columbia.edu

The indirect effect of an exposure on an outcome through an intermediate variable can be identified by a product of regression coefficients under certain causal and regression modeling assumptions. Thus, the null hypothesis of no indirect effect is a composite null hypothesis, as the null holds if either regression coefficient is zero. A consequence is that existing hypothesis tests are either severely underpowered near the origin (i.e., when both coefficients are small with respect to standard errors) or do not preserve type 1 error uniformly over the null hypothesis space. We propose hypothesis tests that (i) preserve level alpha type 1 error, (ii) meaningfully improve power when both true underlying effects are small relative to sample size, and (iii) preserve power when at least one is not. One approach gives a closed-form test that is minimax optimal with respect to local power over the alternative parameter space. Another uses sparse linear programming to produce an approximately optimal test for a Bayes risk criterion. We provide an R package that implements the minimax optimal test.

Statistical methods for improving randomized clinical trial analysis with integrated information from real-world evidencestudies

Shu Yang
North Carolina State University
syang24@ncsu.edu

In this talk, we leverage the complementary features of randomized clinical trials (RCT) and real-world evidence (RWE) to improve treatment effect evaluation. First, we propose calibration weighting estimators that improve the generalizability of the RCT-based estimator by leveraging the representativeness of the RWE study sample. Second, we develop various integrative strategies that borrow RWE to improve HTE (heterogeneity of treatment effect) estimation, while ensuring that any confounding biases present in the RWE do not leak into the proposed estimator. For both objectives, we use the semiparametric efficiency theory to guide the choice of efficient estimators. We apply our proposed methods to estimate the effects of adjuvant chemotherapy in early-stage resected non-small-cell lung cancer integrating data from an RCT and a sample from

the National Cancer Database.

Session 9: Statistics in Biosciences: Recent methodological developments and applications

Generating Survival Times Using Cox Proportional Hazards Models with Cyclic and Piecewise Time-Varying Covariates

♦*Yunda Huang*[1], *Yuanyuan Zhang*[1], *Zong Zhang*[2] *and Peter Gilbert*[1]
[1]Fred Hutchinson Cancer Center
[2]Carnegie Mellon University
Yunda@fredhutch.org

Time-to-event outcomes with cyclic time-varying covariates are frequently encountered in biomedical studies that involve multiple or repeated administrations of an intervention. In this paper, we propose approaches to generating event times for Cox proportional hazards models with both time-invariant covariates and a continuous cyclic and piecewise time-varying covariate. Values of the latter covariate change over time through cycles of interventions and its relationship with hazard differs before and after a threshold within each cycle. The simulations of data are based on inverting the cumulative hazard function and a log link function for relating the hazard function to the covariates. We consider closed-form derivations with the baseline hazard following the exponential, Weibull, or Gompertz distribution. We propose two simulation approaches: one based on simulating survival data under a single-dose regimen first before data are aggregated over multiple-dosing cycles and another based on simulating survival data directly under a multiple-dose regimen. We consider both fixed intervals and varying intervals of the drug administration schedule. The method's validity is assessed in simulation experiments. The results indicate that the proposed procedures perform well in generating data that conform to their cyclic nature and assumptions of the Cox proportional hazards model.

Assessing Treatment Benefit in Immuno-oncology

Marc Buyse
IDDI
marc.buyse@iddi.com

Immuno-oncology is a buoyant field of research, with recently developed drugs showing unprecedented response rates and/or a hope for a meaningful prolongation of the overall survival of some patients. These promising clinical developments have also pointed to the need of adapting statistical methods to best describe and test for treatment effects in randomized clinical trials. We review adaptations to tumor response and progression criteria for immune therapies. Survival may be the endpoint of choice for clinical trials in some tumor types, and the search for surrogate endpoints is likely to continue to try and reduce the duration and size of clinical trials. In situations for which hazards are likely to be non-proportional, weighted logrank tests may be preferred as they have substantially more power to detect late separation of survival curves. Alternatively, there is currently much interest in accelerated failure time models, and in capturing treatment effect by the difference in restricted mean survival times between randomized groups. Finally, generalized pairwise comparisons offer much promise in the field of immuno-oncology, both to detect late emerging treatment effects and as a general approach to personalize treatment choices through a benefit/risk approach.

Statistical and Computational Approaches for the Identification of Novel Viruses and Virus-host Interactions

Fengzhu Sun
University of Southern California
fsun@college.usc.edu

Viruses play important roles in controlling bacterial population size, altering host metabolism, and have broader impacts on the functions of microbial communities, such as human gut, soil, and ocean microbiomes. However, the investigations of viruses and their functions were vastly underdeveloped. Metagenomic studies provide enormous resources for the identifications of novel viruses and their hosts. We developed machine learning based methods, VirFinder and DeepVirFinder, for the identification of novel virus contigs in metagenomic samples. Applications to a liver cirrhosis metagenomic data suggest that viruses play important roles in the development of the disease. We also developed statistical methods, VirHost-Matcher and VirHost-Matcher-Net, for the identification of bacterial hosts of viruses. Applications of these tools to metagenomics data identified a large number of novel virus-host interactions.

A Hybrid Approach for the Strati¡U+FB01¿ed Mark-Speci¡U+FB01¿c Proportional Hazards Model with Missing Covariates and Missing Marks, with Application to Vaccine Ef¡U+FB01¿cacy Trials

♦*Yanqing Sun*[1], *Li Qi*[2], *Fei Heng*[3] *and Peter Gilbert*[4]
[1]The University of North Carolina at Charlotte
[2]Sanofi
[3]University of North Florida
[4]Fred Hutchinson Cancer Research Center and University of Washington)
yasun@uncc.edu

Deployment of the recently licensed CYD-TDV dengue vaccine requires understanding of how the risk of dengue disease in vaccine recipients depends jointly on a host biomarker measured after vaccination (neutralization titer – NAb) and on a "mark" feature of the dengue disease failure event (the amino acid sequence distance of the dengue virus to the dengue sequence represented in the vaccine). The CYD14 phase 3 trial of CYD-TDV measured NAb via case-cohort sampling and the mark in dengue disease failure events, with about a third missing marks. We addressed the question of interest by developing inferential procedures for the strati¡U+FB01¿ed mark-speci¡U+FB01¿c proportional hazards model with missing covariates and missing marks. Two hybrid approaches are investigated that leverage both augmented inverse probability weighting and nearest neighbor hot deck multiple imputation. The two approaches differ in how the imputed marks are pooled in estimation. Our investigation shows that NNHD imputation can lead to biased estimation without properly selected neighborhoods. Simulations show that the developed hybrid methods perform well with unbiased NNHD imputations from proper neighborhood selection. The new methods applied to CYD14 show that NAb is strongly inversely associated with risk of dengue disease in vaccine recipients, more strongly against dengue viruses with shorter distances.

Session 10: New advances in nonparametric and functional data analysis

Low Rank Approximation for Smoothing Spline via Eigensystem Truncation

♦*Yuedong Wang*[1] *and Danqing Xu*[2]
[1]University of California - Santa Barbara
[2]AbbVie
yuedong@pstat.ucsb.edu

Smoothing splines provide a powerful and flexible means for nonparametric estimation and inference. With a cubic time complexity,

fitting smoothing spline models to large data is computationally prohibitive. We use the theoretical optimal eigenspace to derive a low rank approximation of the smoothing spline estimates. We develop a method to approximate the eigensystem when it is unknown and derive error bounds for the approximate estimates. The proposed methods are easy to implement with existing software. Extensive simulations show that the new methods are accurate, fast, and compares favorably against existing methods.

A sparse follow-up procedure for functional contrast tests

Quyen Do and ♦*Pang Du*

Virginia Tech
pangdu@vt.edu

Similar to linear contrast tests for comparing means of multiple treatment groups in the classical ANOVA model, functional contrast tests are designed to compare the mean functions of multiple groups of functional data in a functional ANOVA model. In this talk, we will first examine the extensions of several existing functional ANOVA tests to test on functional contrasts. Then we will introduce a follow-up procedure to identify the region(s) of difference for a significant functional contrast. Simulations will compare the performance of the different versions of functional contrast tests and the follow-up procedure. The analysis of a motivating example on spectral monitoring of hemodialysis will be presented.

Bayesian jackknife empirical likelihood

Yichen Cheng and ♦*Yichuan Zhao*

Georgia State University
yichuan@gsu.edu

Empirical likelihood is a very powerful nonparametric tool that does not require any distributional assumptions. Lazar (2003) showed that in Bayesian inference, if one replaces the usual likelihood with the empirical likelihood, then posterior inference is still valid when the functional of interest is a smooth function of the posterior mean. However, it is not clear whether similar conclusions can be obtained for parameters defined in terms of U-statistics. We propose the so-called Bayesian jackknife empirical likelihood, which replaces the likelihood component with the jackknife empirical likelihood. We show, both theoretically and empirically, the validity of the proposed method as a general tool for Bayesian inference. Empirical analysis shows that the small-sample performance of the proposed method is better than its frequentist counterpart. Analysis of a case-control study for pancreatic cancer is used to illustrate the new approach.

An application of semiparametric method in Nonparametric Regression

Zhijian Li

UCSB
zhijian@pstat.ucsb.edu

Error variance estimation plays a key role in the analysis of homogeneous nonparametric regression models. For a random design model, most methods in the literature for error variance estimation assume the independence between the predictor variable and the random errors. Here we derive the optimal semiparametric efficiency bound for the error variance without such an independence assumption. Using the semiparametric method, we can also propose a residual-based efficient estimator and establish its asymptotic normality.

Session 11: Recent advances in statistical methods for complex biomedical data

Decomposing cell compositional effect in bulk tissue gene expression analysis

♦*Shaoyu Li*[1]*, Xue Wang*[2]*, Duan Chen*[1] *and Jinjing Zhang*[1]
[1]University of North Carolina
[2]Mayo Clinics
sli23@uncc.edu

Since different distributions of cellular compositions in bulk tissue samples may confound the overall changes in gene expression between groups of individuals, the correction of cell compositional effect in bulk tissue gene expression data analysis is essential to reduce false positive rates and identify reliable genes that are directly affected by disease status and, therefore, potential drug targets. In this work we propose a statistical framework to decompose the overall changes in average gene expression between two groups of individuals to cell composition effect (CCE) and direct transcriptional effect (DTE) that is due to transcriptional mechanisms. In particular, we derive a propensity score weighted estimator for the decomposition and a permutation-based procedure to test the statistical significance. We conduct comprehensive simulations to illustrate the merits of our proposed method. We demonstrate that it is promising to use our method to detect genes that are directly associated with type II diabetes.

Robust Estimation of Heterogeneous Treatment Effects using Electronic Health Record Data

Ruohong Li[1]*,* ♦*Honglang Wang*[2] *and Wanzhu Tu*[1]
[1]Indiana University
[2]Indiana University-Purdue University Indianapolis
hlwang@iupui.edu

Estimation of heterogeneous treatment effects is an essential component of precision medicine. Model and algorithm-based methods have been developed within the causal inference framework to achieve valid estimation and inference. Existing methods such as the A-learner, R-learner, modified covariates method (with and without efficiency augmentation), inverse propensity score weighting, and augmented inverse propensity score weighting have been proposed mostly under the square error loss function. The performance of these methods in the presence of data irregularity and high dimensionality, such as that encountered in electronic health record (EHR) data analysis, has been less studied. In this research, we describe a general formulation that unifies many of the existing learners through a common score function. The new formulation allows the incorporation of least absolute deviation (LAD) regression and dimension reduction techniques to counter the challenges in EHR data analysis. We show that under a set of mild regularity conditions, the resultant estimator has an asymptotic normal distribution. Within this framework, we proposed two specific estimators for EHR analysis based on weighted LAD with penalties for sparsity and smoothness simultaneously. Our simulation studies show that the proposed methods are more robust to outliers under various circumstances. We use these methods to assess the blood pressure-lowering effects of two commonly used antihypertensive therapies.

A Fast and Efficient Likelihood Approach for Genome-wide Mediation Analysis

Janaka Liyanage[1]*, Jeremie Estepp*[1]*, Kumar Srivastava*[1]*, Yun Li*[2]*, Motomi Mori*[1] *and* ♦*Guolian Kang*[1]
[1]St. Jude Children's Research Hospital
[2]The University of North Carolina at Chapel Hill
Guolian.Kang@stjude.org

Due to many advantages such as higher power of detecting the association of rare genetic variants in human disorders and cost effectiveness of the study design, extreme phenotype sequencing (EPS) is a rapidly emerging study design in epidemiological and clinical studies investigating how genetic variations associate with underlying disease mechanisms. However, investigation of the mediation effect of genetic variants in underlying disease mechanisms is strictly restrictive under the EPS design because existing methods cannot well accommodate the non-random extreme tails sampling process incurred by the EPS design. In this paper, we propose a likelihood approach for testing the mediation effect of genetic variants through continuous and binary mediators on a continuous phenotype under the EPS design (GMEPS). Besides implementing in EPS design, it also can be utilized as a general mediation analysis procedure. Extensive simulations and two real data applications of a genome-wide association study of benign ethnic neutropenia under EPS design and a candidate-gene study of neurocognitive performance in patients with sickle cell disease under random sampling design demonstrate the superiority of the GMEPS under the EPS design over widely used mediation analysis procedures, while demonstrating compatible capabilities under the general random sampling framework.

Multi-marker survival tests for interval censored data

♦*Chenxi Li*[1]*, Di Wu*[1] *and Qing Lu*[2]
[1]Michigan State University
[2]University of Florida
cli@msu.edu

The development of set-based genetic-survival association tests has been focusing on right-censored survival outcomes. However, interval-censored failure time data arise widely from health science studies, especially those on the development of chronic diseases. In this paper, we proposed a suite of set-based genetic association and interaction tests for interval-censored survival outcomes under a unified weighted-V-statistic framework. Besides dealing with interval censoring, the new tests can account for genetic effect heterogeneity and accommodate left truncation of survival outcomes. Simulation studies showed that the new tests perform well in terms of size and power under various scenarios and that the new interaction test is more powerful than the standard likelihood ratio test for testing gene-gene/gene-environment interactions. The practical utility of the developed tests was illustrated by a genome-wide association study of age to early childhood caries.

Session 12: Recent Advances in Causal Inference With Applications to Public Health and Policy

Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment

Kosuke Imai[1]*,* ♦*Zhichao Jiang*[2]*, James Greiner*[3]*, Ryan Halen*[3] *and Sooahn Shin*[1]
[1]Harvard University
[2]University of Massachusetts Amherst
[3]Harvard Law School
zhichaojiang@umass.edu

Despite an increasing reliance on fully-automated algorithmic decision making in our day-to- day lives, human beings still make highly consequential decisions. As frequently seen in business, healthcare, and public policy, recommendations produced by algorithms are provided to human decision-makers in order to guide their decisions. While there exists a fast-growing literature evaluating the bias and fairness of such algorithmic recommendations, an overlooked question is whether they help humans make better decisions. We develop a statistical methodology for experimentally evaluating the causal impacts of algorithmic recommendations on human deci- sions. We also show how to examine whether algorithmic recommendations improve the fairness of human decisions and derive the optimal decision rules under various settings. We apply the proposed methodology to the preliminary data from the first-ever randomized controlled trial that evaluates the pretrial Public Safety Assessment (PSA) in the criminal justice system. A goal of the PSA is to help judges decide which arrested individuals should be released. On the basis of the preliminary data available, we find that providing the PSA to the judge has little overall impact on the judge's decisions and subsequent arrestee behavior. However, our analysis provides some potentially suggestive evidence that the PSA may help avoid unnecessarily harsh decisions for female arrestees regardless of their risk levels while it encourages the judge to make stricter decisions for male arrestees who are deemed to be risky.

Propensity score weighting analysis of survival outcomes using pseudo-observations

Shuxi Zeng[1]*, Fan Li*[1]*, Liangyuan Hu*[2] *and* ♦*Fan Li*[3]
[1]Duke University
[2]Rutgers University
[3]Yale University
fan.f.li@yale.edu

Survival outcomes are common in comparative effectiveness studies and require unique handling because they are usually incompletely observed due to right-censoring. A "once for all" approach for causal inference with survival outcomes constructs pseudo-observations and allows standard methods such as propensity score weighting to proceed as if the outcomes are completely observed. We propose a general class of model-free causal estimands with survival outcomes on user-specified target populations. We develop corresponding propensity score weighting estimators based on the pseudo-observations and establish their asymptotic properties. In particular, utilizing the functional delta-method and the von Mises expansion, we derive a new closed-form variance of the weighting estimator that takes into account the uncertainty due to both pseudo-observation calculation and propensity score estimation. This allows valid and computationally efficient inference without resampling. We also prove the optimal efficiency property of the overlap weights within the class of balancing weights for survival outcomes. The proposed methods are applicable to both binary and multiple treatments. Extensive simulations are conducted to explore the operating characteristics of the proposed method versus other commonly used alternatives. We apply the proposed method to compare the causal effects of three popular treatment approaches for prostate cancer patients.

A negative correlation strategy for bracketing in difference-in-differences

♦*Ting Ye*[1]*, Luke Keele*[1]*, Raiden Hasegawa*[2] *and Dylan Small*[1]
[1]University of Pennsylvania
[2]Google
tingye@wharton.upenn.edu

The method of difference-in-differences (DID) is widely used to study the causal effect of policy interventions in observational studies. DID employs a before and after comparison of the treated and control units to remove bias due to time-invariant unmeasured con-

founders under the parallel trends assumption. Estimates from DID, however, will be biased if the outcomes for the treated and control units evolve differently in the absence of treatment, namely if the parallel trends assumption is violated. We propose a general identification strategy that leverages two groups of control units whose outcomes relative to the treated units exhibit a negative correlation, and achieves partial identification of the average treatment effect for the treated. The identified set is of a union bounds form that involves the minimum and maximum operators, which makes the canonical bootstrap generally inconsistent and naive methods overly conservative. By utilizing the directional inconsistency of the bootstrap distribution, we develop a novel bootstrap method to construct confidence intervals for the identified set and parameter of interest when the identified set is of a union bounds form, and we theoretically establish the uniform asymptotic validity of the proposed method. We develop a simple falsification test and sensitivity analysis. We apply the proposed strategy for bracketing to study whether minimum wage laws affect employment levels.

Session 13: New Statistical Methods for Competing Risks

Doubly Robust Estimation of the Hazard Difference for Competing Risks Data

♦*Denise Rava and Ronghui (Lily) Xu*
UC San Diego
rxu@health.ucsd.edu

We consider the conditional treatment effect for competing risks data in observational studies. While it is described as a constant difference between the hazard functions given the covariates, we do not assume the additive hazards model in order to adjust for the covariates. We derive the efficient score for the treatment effect using modern semiparametric theory, as well as two doubly robust scores with respect to both the assumed propensity score for treatment and the censoring model, and the outcome models for the competing risks. We provide the asymptotic distributions of the estimators when the two sets of working models are both correct, or when only one of them is correct. We study the inference based on these estimators using simulation. The estimators are applied to the data from a cohort of Japanese men in Hawaii followed since 1960s in order to study the effect of midlife drinking behavior on late life cognitive outcomes.

Variable selection in semiparametric transformation regression with interval-censored competing risks data

♦*Fatemeh Mahmoudi and Xuewen Lu*
University of Calgary
fatemeh.mahmoudi@ucalgary.ca

In the framework of variable selection problems, penalized regression is a popular approach. Although there have been numerous researches on penalized variable selection methods for standard time-to-event data and regression models, they are not applicable when data are interval-censored competing risks. In this context, we develop a penalized variable selection procedure that is able to handle such data in a broad class of semiparametric transformation regression models, which contain some popular models such as the proportional and non-proportional hazards models as special cases and allow for direct assessment of covariate effects on the cumulative incidence or sub-distribution function of competing risks. The proposed penalized variable selection strategy can simultaneously handle variable selection and parameter estimation. We rigorously establish the asymptotic properties of the proposed penalized estimators and modify the EM algorithm and coordinate descent algorithm for implementation. Simulation studies are conducted to demonstrate the good performance of the proposed method. Some real data examples are used for illustration.

Semiparametric Marginal Regression for Clustered Competing Risks Data with Missing Cause of Failure

Wenxian Zhou[1], ♦*Giorgos Bakoyannis*[1], *Ying Zhang*[2] *and Constantin Yiannoutsos*[1]
[1]Indiana University
[2]University of Nebraska Medical Center
gbakogia@iu.edu

Clustered competing risks data are commonly encountered in multicenter studies. The analysis of such data is often complicated due to informative cluster size, a situation where the outcomes under study are associated with the size of the cluster. In addition, cause of failure is frequently incompletely observed in real-world settings. To the best of our knowledge, there is no methodology for population-averaged analysis with clustered competing risks data with informative cluster size and missing causes of failure. To address this problem, we consider the semiparametric marginal proportional cause-specific hazards model and propose a maximum partial pseudolikelihood estimator under a missing at random assumption. To make the latter assumption more plausible in practice, we allow for auxiliary variables that may be related to the probability of missingness. The proposed method does not impose assumptions regarding the within-cluster dependence and allows for informative cluster size. The asymptotic properties of the proposed estimators for both regression coefficients and infinite-dimensional parameters, such as the marginal cumulative incidence functions, are rigorously established. Simulation studies show that the proposed method performs well and that methods that ignore the within-cluster dependence and the informative cluster size lead to invalid inferences. The proposed method is applied to competing risks data from a large multicenter HIV study in sub-Saharan Africa where a significant portion of causes of failure is missing.

Joint Correlation Learning for Semi-competing Risks Outcomes with Ultrahigh Dimensional Covariates

Mengjiao Peng[1] *and* ♦*Liming Xiang*[2]
[1]East China Normal University, Shanghai
[2]Nanyang Technological University, Singapore
lmxiang@ntu.edu.sg

Ultrahigh-dimensional gene features are often collected in modern cancer studies in which the number of gene features p is extremely larger than sample size n. While gene expression patterns have been shown to be related to patients' survival in microarray-based gene expression studies, one has to deal with the challenges of ultrahigh-dimensional genetic predictors for survival prediction and genetic understanding of the disease in precision medicine. The problem becomes more complicated when two types of survival endpoints, distant metastasis-free survival (DMFS) and overall survival (OS), are of interest in the study and outcome data can be subject to semi-competing risks due to the fact that DMFS is possibly censored by OS but not vice versa. Our focus in this talk is to extract important features, which have great impacts on both DMFS and OS jointly, from massive gene expression data in the semi-competing risks setting. We propose a model-free screening method based on the ranking of the correlation between gene features and the joint survival function of two endpoints. The method accounts for the relationship between two endpoints in a simply defined utility measure that is easy to understand and calculate. We show its favorable theoretical properties and evaluate its finite sample performance through

extensive simulation studies. Finally, an application to classifying breast cancer data demonstrates the utility of the proposed method in practice.

Session 14: Recent Advancements in Large-Scale and High-dimensional Inference

Individual data protected meta-analysis of large scale healthcare data from multiple sites

Tianxi Cai[1], ♦*Molei Liu*[1] *and Yin Xia*[2]
[1]Harvard Chan School of Public Health
[2]Fudan University
molei_liu@g.harvard.edu

Evidence based decision making often relies on meta-analyzing multiple studies, which enables more precise estimation, more powerful knowledge extraction, and investigation of generalizability. For large-scale healthcare data from multiple sites, such integration usually encounters practical problems including individual data privacy, high dimensionality, and heterogeneity. I will introduce our recent progress in developing individual data-protected meta-analysis methods for electronic health records (EHR) data from multiple healthcare systems. Our methods can be used for model-based estimation, prediction, variable selection, and hypothesis testing. They accommodate ultra-high dimensionality and heterogeneity of the data across different sites and are proved to be statistically efficient. We also demonstrate their use through real-world examples.

Tuning-free large-scale multivariate regression via shrinkage estimation

Yihe Wang[1] *and* ♦*Dave Zhao*[2]
[1]Facebook, Inc.
[2]University of Illinois Urbana-Champaign
sdzhao@illinois.edu

We propose a new method for multivariate linear regression problems where the number of features is less than the sample size but the number of outcomes is extremely large. Estimation is typically performed using penalized regression estimators like the group lasso, but these require parameter tuning that is computationally untenable in these large-scale problems. We take a different approach, motivated by ideas from shrinkage estimation and compound decision theory, that performs linear shrinkage on ordinary least squares parameter estimates. Our approach is extremely computationally efficient and tuning-free, and we show that it asymptotically outperforms the ordinary least squares estimator without any structural assumptions on the true regression coefficients. In simulations and an analysis of single-cell RNA-seq data, our method outperforms the group lasso and other penalized procedures.

LinDA: Linear Models for Differential Abundance Analysis of Microbiome Compositional Data

♦*Huijuan Zhou*[1], *Xianyang Zhang*[2], *Kejun He*[3] *and Jun Chen*[4]
[1]Texas A&M University; Renmin University of China
[2]Texas A&M University
[3]Renmin University of China
[4]Mayo Clinic
huijuan@stat.tamu.edu

One fundamental statistical task in microbiome data analysis is differential abundance analysis, which aims to identify microbial taxa whose abundance covaries with a variable of interest. Although the main interest is on the change in 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 abundances 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. We implemented the proposed method in the R package LinDA (https://github.com/zhouhj1994/LinDA).

A flexible model-free prediction-based framework for feature ranking

♦*Jingyi Jessica Li*[1], *Yiling Chen*[1] *and Xin Tong*[2]

[1]University of California Los Angeles

[2]University of Southern California

jli@stat.ucla.edu

Despite the availability of numerous statistical and machine learning tools for joint feature modeling, many scientists investigate features marginally, i.e., one feature at a time. This is partly due to training and convention but also roots in scientists' strong interests in simple visualization and interpretability. As such, marginal feature ranking for some predictive tasks, e.g., prediction of cancer driver genes, is widely practiced in the process of scientific discoveries. In this work, we focus on marginal ranking for binary classification, one of the most common predictive tasks. We argue that the most widely used marginal ranking criteria, including the Pearson correlation, the two-sample t test, and two-sample Wilcoxon rank-sum test, do not fully take feature distributions and prediction objectives into account. To address this gap in practice, we propose two ranking criteria corresponding to two prediction objectives: the classical criterion (CC) and the Neyman-Pearson criterion (NPC), both of which use model-free nonparametric implementations to accommodate diverse feature distributions. Theoretically, we show that under regularity conditions, both criteria achieve sample-level ranking that is consistent with their population-level counterpart with high probability. Moreover, NPC is robust to sampling bias when the two class proportions in a sample deviate from those in the population. This property endows NPC good potential in biomedical research where sampling biases are ubiquitous. We demonstrate the use and relative advantages of CC and NPC in simulation and real data studies. Our model-free objective-based ranking idea is extendable to ranking feature subsets and generalizable to other prediction tasks and learning objectives.

Session 15: Statistical Learning and Variable Selection

A Tweedie Compound Poisson Model in RKHS

Yi Yang
McGill University
yi.yang6@mcgill.ca

Abstract Tweedie models have wide applications in natural science, healthcare research, actuarial science, etc. The performance of existing Tweedie models can be limited by today's complex data problems with challenging characteristics such as high-dimensionality, nonlinear effects, high-order interactions. Motivated by these challenges, we propose a kernel Tweedie model with integrated variable selection. The non parametric nature of the the proposed method along with the abundance of available kernel functions provides much needed modeling flexibility and capability. The resulting sparsity due to the variable selection also improves the interpretability and the prediction accuracy. We perform extensive simulation studies to justify the prediction and variable selection accuracy of our method, and demonstrate the applications in ratemaking and loss-reserving in general insurance. The model is implemented in an efficient and user-friendly R package.

Simultaneous Variable Selection and Covariance Estimation in High Dimensional Linear Mixed Model

Xin Liu
Shanghai University of Finance and Economics
liu.xin@mail.shufe.edu.cn

The linear mixed effect model has been widely used to analyze complex data invarious fields. However, an essential research gap still remains on how to select both the fixed and random effects simultaneously, especially for longitudinal data. In this paper, we proposed an iterative procedure for simultaneous variable selection for both fixed and random effects as well as estimation of the corresponding covariance matrix for linear mixed effect model (LMM) for longitudinal data in high dimension setting. The procedure employs penalized methods for fixed effect selection, and graphical methods to estimate random effects by screening the row (or column) of the covariance matrix. Asymptotically, the proposed procedure is proved to embrace selection and estimation accuracy. Numerically, both the simulation data and a real data set, 2016 American National Election Survey (ANES), are conducted to examine the performance of this procedure. The proposed procedure can easily be extended for other complex data.

Nonconvex clustering with random projection

Xiaodong Yan
Shandong University
yanxiaodong@sdu.edu.cn

Clustering is one preliminary operation of data processing in data mining and statistics. Novelly introduced convex or concave fusion clustering method outperforms classical tools such as k-means and hierarchical clustering in terms of stability in computation, simultaneously identifying cluster center and number and characterizing attainment of global or local optimality. However, this newly relaxed alternative is beset by drastically burdensome complexity of fusion, causing intensive computation. This article takes this tough issue and proposes random projection ADMM algorithm based on binary distribution and develops double random projection ADMM for high-dimensional clustering issue. Both new formulations largely outperform classical ADMM algorithm because they intensively accelerate calculative speed through reducing the computational complexity and improve the clustering accuracy through adopting multiple random projections under a new proposed evaluation criterion. We demonstrate the convergence property of this new algorithm and check the performance of random projection clustering procedure on both simulated and real data examples.

Bootstrap Inference for the Finite Population Mean under Complex Sampling Designs

♦*Zhonglei Wang*[1]*, Liuhua Peng*[2] *and Jae-Kwang Kim*[3]
[1]Xiamen University
[2]The University of Melbourne
[3]Iowa State University
wangzl@xmu.edu.cn

Bootstrap is a useful computational tool for making statistical inference, but the conventional bootstrap may lead to erroneous analysis results under complex survey sampling. How to properly develop bootstrap methods under complex sampling is a fundamental research problem that is less explored in the literature. In this paper, we propose a unified bootstrap method for stratified multi-stage cluster sampling as well as some commonly used single-stage sampling designs, including Poisson sampling, simple random sampling, probability proportional to size sampling with replacement. In the proposed bootstrap, studentization is used to make more accurate statistical inference, and the same sampling design is applied to generate bootstrap samples from the associated bootstrap finite populations. Second-order accuracy of the proposed bootstrap method is established by the Edgeworth expansion. Simulation studies confirm that the proposed bootstrap method outperforms the commonly used Wald-type method in terms of coverage rate, especially when the sample size is not large.

Session 16: Recent advances in treatment evaluation and risk prediction with survival data

Dynamic risk prediction of adverse outcomes of ICU patients with sepsis using machine learning methods

Chung-Chou H. Chang
University of Pittsburgh
changj@pitt.edu

We used machine learning approaches to predict the probabilities of readmission or death among sepsis patients admitted to an intensive care units (ICUs) based on demographics information and daily measured lab values. The machine learning approaches integrated the landmarking method to handle longitudinal trajectories of repeated measurements more efficiently. Data used in this analysis were collected from a health care system in western Pennsylvania. Prediction performance was assessed by the time-dependent Brier score and time-dependent c-statistic.

On restricted mean time in favor of treatment

Lu Mao
N/A
lmao@biostat.wisc.edu

The restricted mean time in favor (RMT-IF) of treatment is a nonparametric effect size for complex life history data. The estimand is defined as the net average time the treated spend in a more favorable state than the untreated as opposed to vice versa over a fixed time window. It generalizes the familiar restricted mean survival time from the two-state life-death model to account for possible intermediate stages in disease progression. The overall estimand admits an elegant decomposition into stage-wise effects, with the standard restricted mean survival time as a component. Alternate expressions of the overall and stage-wise estimands as integrals of

the marginal survival functions for a sequence of landmark transitioning events facilitate their estimation by simple plug-in Kaplan–Meier estimators.The dynamic profile of the estimated treatment effects as a function of follow-up time can be visualized using a multilayer, cone-shaped "bouquet plot". Simulation studies under realistic settings show that the RMT-IF approach provides meaningful and accurate quantification of the treatment effect and outperforms traditional tests on time to the first event thanks to its fuller utilization of patient data. We illustrate the proposed methods on a colon cancer trial with relapse and death as outcomes and a cardiovascular trial with recurrent hospitalizations and death as outcomes.

The benefit-risk assessment of new treatments using generalized pairwise comparisons, a population- and individual-level perspective

julien peron

hospices de civils de lyon

julien.peron@chu-lyon.fr

A novel statistical approach to the analysis of randomized clinical trials uses all pairwise comparisons between two patients, one in the treatment arm and one in the control arm. Each pair favors treatment (""win""), control (""loss""), or neither. The ""net treatment benefit"" is the difference between the proportion of wins minus the proportion of losses. Pairwise comparisons can incorporate several outcomes of interest and several thresholds of clinical relevance in the analysis, and as such, they can be used to personalize treatment choices and to assess the benefit/risk of randomized therapeutic interventions in a rigorous yet flexible manner (Buyse 2010). The advantages and limitations of generalized pairwise comparisons will be illustrated using two typical examples. For a single time-to-event endpoint, the net survival benefit is a meaningful measure of treatment effect whether or not hazards are proportional. When a delayed treatment effect is anticipated, for example in immune-oncology trials, the net benefit is appealing because it stresses benefits that are clinically worthwhile on the time scale. The test based on the net survival benefit can also gain power as compared to the traditional logrank test if interest focuses on long-term survival differences (Péron 2016a). Most anticancer treatment have substantial toxicities that may counterbalance treatment benefits. Generalized pairwise comparisons can be used to assess the benefit-risk balance of new treatments. This will be illustrated using several randomized trials in patients with metastatic pancreatic cancer (Péron 2016b). When multiple endpoints are analyzed simultaneously, generalized pairwise comparisons use a single set of endpoint priorities and can use thresholds of minimal clinical relevance for all endpoints. The set of priorities and thresholds can be defined at the population-level, or at the individual level, in order to assess the individual Net Treatment Benefit. References Buyse M. Generalized pairwise comparisons for prioritized outcomes in the two-sample problem. Stat Med 2010; 29: 3245-3257. Péron J, Roy P, Ozenne B, Roche L, Buyse M. The net chance of a longer survival as a patient-oriented measure of benefit in randomized clinical trials. JAMA Oncology 2016a; 2:901-5. Péron J, Roy P, Conroy T, Desseigne F, Ychou M, Gourgou-Bourgade S, Stanbury T, Roche L, Ozenne B, Buyse M. An assessment of the benefit-risk balance of FOLFIRINOX in metastatic pancreatic adenocarcinoma. Oncotarget 2016b;7:82953-82960.

Session 17: Statistics in Microbiome Research

Microbiome multi-omics integration using compositional reduced rank regression for relative abundance data

Chenxiao Hu and ♦Duo Jiang

Oregon State University

jiangd@stat.oregonstate.edu

The emergence of microbiome multi-omics studies calls for effective statistical methods to infer the associations between microbiome measures and another omics data type (e.g. metabolomics). Regression analysis for integrating multi-omics data type is challenging due to the high-dimensional nature of both data types. In regression models where both the response and the predictor data are high-dimensional, reduced rank regression (RRR) is often a useful tool because it reduces the number of parameters to be estimated by imposing a low rank structure on the coefficient matrix. However, existing RRR methods fail to account for the compositionality of microbiome data. Specifically, using microbiome sequencing, the absolute abundances of microbes are unobservable and microbiome composition is only characterized by the relative abundance (RA) of a microbe relative to other microbes. To resolve this challenge, we propose a new multivariate regression method to estimate the effects of microbiome absolute abundances which necessitates RA data only. Our model explicitly incorporates the unknown total microbial abundance into a reduced rank regression model with nuclear penalty. An ADMM-based algorithm is developed to estimate the microbiome-response association matrix. Simulation studies demonstrate that the proposed approach has superior performance compared to the standard RRR model in terms of both estimation precision and prediction accuracy.

Measurement error in compositional data and the replicability of microbiome studies

Amy Willis

University of Washington

adwillis@uw.edu

The composition of bacterial taxa in a microbiome is an important parameter to estimate given the critical role that microbiomes play in human and environmental health. However, high throughput sequencing distorts the true composition of microbial communities. Sequencing mock communities – artificially constructed microbiomes of known composition – clearly illustrates that observed composition is a biased estimate of true composition, with certain taxa consistently overobserved or underobserved compared to their true relative abundance. We propose a statistical model for bias in compositional data, illustrating its performance on data from the Vaginal Microbiome Consortium. We also present our assessment of the replicability of human microbiome data, where we find that different sequencing facilities find substantially different signals in identical specimens. We conclude with recommendations for the design and analysis of microbiome studies.

Dimension Reduction of Longitudinal Microbiome Data via Tensor Functional SVD

♦*Pixu Shi, Rungang Han and Anru Zhang*

Duke University

pixu.shi@duke.edu

The reduction of sequencing cost has prompted more microbiome studies with longitudinal measurements of bacterial abundance. Longitudinal microbiome data can often be formatted into a high-dimensional order-3 tensor with three modes representing the subject, time, and bacteria respectively. Since the time of measurement for different subjects can be highly variable, the values of

such the order-3 tensor are typically not well-aligned, making it challenging to analyze the trajectory of bacterial abundance over time and identify key bacteria associated with time or clinical phenotypes. In this paper, we propose a new tensor functional SVD method that performs dimension reduction to assist the analysis of high-dimensional longitudinal microbiome data. The new method can extract the key components in the trajectories of bacterial abundance, identify representative bacterial taxa for these key trajectories, and group subjects based on the change of bacteria abundance over time. The new method is also flexible to handle microbiome measurements at irregular time points for different subjects.

Joint modeling of zero-inflated longitudinal proportions and time-to-event data with application to a gut microbiome study

♦*Jiyuan Hu*[1]*, Chan Wang*[1]*, Martin Blaser*[2] *and Huilin Li*[1]
[1]NYU Grossman School of Medicine
[2]Rutgers University
jiyuan.hu@nyumc.org

Recent studies have suggested that the temporal dynamics of the human microbiome may have associations with human health and disease. An increasing number of longitudinal microbiome studies, which record time to disease onset, aim to identify candidate microbes as biomarkers for prognosis. Owing to the ultra-skewness and sparsity of microbiome proportion (relative abundance) data, directly applying traditional statistical methods may result in substantial power loss or spurious inferences. We propose a novel joint modeling framework [JointMM], which is comprised of two submodels: a longitudinal sub-model called zero-inflated scaled-Beta generalized linear mixed-effects regression to depict the temporal structure of microbial proportions among subjects; and a survival sub-model to characterize the occurrence of an event and its relationship with the longitudinal microbiome proportions. JointMM is specifically designed to handle the zero-inflated and highly skewed longitudinal microbial proportion data and examine whether the temporal pattern of microbial presence and/or the non-zero microbial proportions are associated with differences in the time to an event. The longitudinal sub-model of JointMM also provides the capacity to investigate how the (time-varying) covariates are related to the temporal microbial presence/absence patterns and/or the changing trend in non-zero proportions. Comprehensive simulations and real data analyses are used to assess the statistical efficiency and interpretability of JointMM.

Session 18: Statistical Text Mining

How Much Can Machines Learn Finance From Chinese Text Data?

♦*Jianqing Fan, Lirong Xue and Yang Zhou*
Princeton University
jqfan@princeton.edu

Most studies on equity markets using text data focus on English-based specified sentiment dictionaries or topic modeling. However, can we predict the impact of news directly from the text data? How much can we learn from such a direct approach? We present here a new framework for learning text data based on the factor model and sparsity regularization, called FarmPredict, to let machines learn financial returns automatically. Unlike other dictionary-based or topic models that have stringent pre-screening processes, our framework allows the model to extract information more fully from the whole article. We demonstrate our study on the Chinese stock market, as Chinese text has no natural spaces between words and phrases and the Chinese market has a very large proportion of retail investors. These two specific features of our study differ significantly from the previous literature that focuses on English-text and the U.S. market. We validate our method using the literature on the Chinese stock market with several existing approaches. We show that positive sentiments scored by our FarmPredict approach generate on average 83 bps stock daily excess returns, while negative news has an adverse impact of 26 bps on the days of news announcements, where both effects can last for a few days. This asymmetric effect aligns well with the short-sale constraints in the Chinese equity market. As a result, we show that the machine-learned sentiments do provide sizable predictive power with an annualized return of 116% with a simple investment strategy and the portfolios based on our model significantly outperform other models. This lends further support that our FarmPredict can learn the sentiments embedded in financial news. Our study also demonstrates the far-reaching potential of using machines to learn text data.

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

♦*Florentina Bunea, Mike Bing, Marten Wegkamp and Seth Strimas-Mackey*
Cornell University
fb238@cornell.edu

This talk regards the estimation of high-dimensional, discrete, possibly sparse, mixture models in the context 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 also 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 $\widehat{A}$ of A. For any estimator $\widehat{A}$ that satisfies carefully detailed conditions for proximity to A, we show that the resulting estimator of T retains the properties established for the MLE. Our theoretical results allow the ambient dimensions K and p to grow with the sample sizes. Our main application is to the estimation of 1-Wasserstein distances between document generating distributions. We propose, estimate and analyze new 1-Wasserstein distances between alternative probabilistic document representations, at the word and topic level, respectively. We derive finite sample bounds on the estimated proposed 1-Wasserstein distances. For word level document-distances, we provide contrast with existing rates on the 1-Wasserstein distance between standard empirical frequency estimates. The effectiveness of the proposed 1-Wasserstein distances is illustrated by an analysis of an IMDb movie reviews data set.

Topic Analysis of Statistical Abstracts

Pengsheng Ji[1]*, Jiashun Jin*[2]*,* ♦*Tracy Ke*[3] *and Wanshan Li*[2]

[1]University of Georgia
[2]Carnegie Mellon University
[3]Harvard University
zke@fas.harvard.edu
We have collected and cleaned a data set consisting of the bibtex of 83K papers published in 36 journals in statistics and related fields, spanning 41 years. The data set motivates many interesting research problems on the publications of statisticians. One problem of great interest is understanding the topic interests in these publications. We use the abstracts of statistical papers to identify 11 representative topics by the topic modeling algorithm in Ke and Wang (2017). Based on the topic model results, we study the topic interests of journals (e.g., "friendliest" journal for each topic), topic interests of individual authors, the cross-topic citations, topic ranking, and how the topic of a paper affects its citations. Our studies use several new models and novel approaches, and require substantial efforts in modeling, setting key parameters, labeling discovered groups, and interpreting the results.

Session 19: Recent Developments on Network Data Analysis

Optimal adaptivity of Signed-Polygon for network testing
Jiashun Jin
Carnegie Mellon University
jiashun@cmu.edu
Given a symmetric social network, we are interested in testing whether it has only one community or multiple communities. The desired tests should (a) accommodate severe degree heterogeneity, (b) accommodate mixed-memberships, (c) have a tractable null distribution, and (d) adapt automatically to different levels of sparsity, and achieve the optimal phase diagram. How to find such a test is a challenging problem. We propose the Signed Polygon as a class of new tests, including the Signed Triangle (SgnT) and the Signed Quadrilateral (SgnQ) are special cases. We show that both the SgnT and SgnQ tests satisfy (a)-(d), and especially, they work well for both very sparse and less sparse networks. Our approach compares favorable with existing tests (which do not necessary satisfy (a)-(d)). Our tests are useful in measuring co-authorship diversity and in setting stopping rules for recursive community detection algorithms. For example, by using a large-scale data set on statisticians' publication data set which we collected and cleaned and by combining our tests with the SCORE algorithm (Jin, 2015), we have obtained a community co-authorship tree for statisticians.

Fast Network Community Detection with Profile-Pseudo Likelihood Methods
Ji Zhu
University of Michigan
jizhu@umich.edu
The stochastic block model is one of the most studied network models for community detection. It is well-known that most algorithms proposed for fitting the stochastic block model likelihood function cannot scale to large-scale networks. One prominent work that overcomes this computational challenge is Amini et al. (2013), which proposed a fast pseudo-likelihood approach for fitting stochastic block models to large sparse networks. However, this approach does not have convergence guarantee, and is not well suited for small- or medium- scale networks. In this article, we propose a novel likelihood based approach that decouples row and column labels in the likelihood function, which enables a fast alternating maximization; the new method is computationally efficient, performs well for both small and large scale networks, and has provable convergence guarantee. We show that our method provides strongly consistent estimates of the communities in a stochastic block model. As demonstrated in simulation studies, the proposed method outperforms the pseudo-likelihood approach in terms of both estimation accuracy and computation efficiency, especially for large sparse networks. We further consider extensions of our proposed method to handle networks with degree heterogeneity and bipartite properties. This is joint work with Jiangzhou Wang, Jingfei Zhang, Binghui Liu, and Jianhua Guo.

Testing correlation of unlabeled random graphs
Yihong Wu[1], ♦*Jiaming Xu*[2] *and Sophie H. Yu*[2]
[1]Yale University
[2]Duke University
jiaming.xu868@duke.edu
The problem of detecting the edge correlation between two random graphs with unlabeled nodes is studied. This is formalized as a hypothesis testing problem, where under the null hypothesis, the two graphs are independently generated; under the alternative, the two graphs are edge-correlated under some latent node correspondence but have the same marginal distributions as the null. For both Gaussian-weighted complete graphs and dense ER graphs, we determine the sharp threshold at which the optimal testing error probability exhibits a phase transition from zero to one. For sparse ER graphs, we determine the threshold within a constant factor. The proof of the impossibility results is an application of the conditional second-moment method, where we bound the truncated second moment of the likelihood ratio by carefully conditioning on the typical behavior of the intersection graph (consisting of edges in both observed graphs) and taking into account the cycle structure of the induced random permutation on the edges.

Recommendation system with social network data by tensor factorization
Yiyuan Liu, Ya Wang and ♦*Bingyi Jing*
HKUST
majing@ust.hk
We consider recommendations for some popular apps, e.g., Foursquare and Gowalla. Such location-based social networks have some special characteristics, e.g., sparse check-in data, long-tail distribution, user's check-in behavior. In this talk, we propose a tensor-based method which are tailored to the above characteristics. Firstly, a user-POIs-time tensor is used to model all user's check-in behaviors. Secondly, log transformation is used to eliminate the influence of long-tail distribution. Third social information is fused into a tensor factorization framework. Finally, collaborative filtering is used to fill in the missing entries of a tensor after clustering the user mode. Experiments show that the recommendations can be drastically improved by taking these measures.

Session 20: Causal inference methods: propensity score, matching, imputation and more

Propensity score analysis methods with balancing constraints: a Monte Carlo study
Yan Li and ♦*Liang Li*
MD Anderson Cancer Center
LLi15@mdanderson.org
The inverse probability weighting is an important propensity score weighting method to estimate the average treatment effect. Re-

cent literature shows that it can be easily combined with covariate balancing constraints to reduce the detrimental effects of excessively large weights and improve balance. Other methods are available to derive weights that balance covariate distributions between the treatment groups without the involvement of propensity scores. We conducted comprehensive Monte Carlo experiments to study whether the use of covariate balancing constraints circumvent the need for correct propensity score model specification, and whether the use of a propensity score model further improves the estimation performance among methods that use similar covariate balancing constraints. We compared simple inverse probability weighting, two propensity score weighting methods with balancing constraints (covariate balancing propensity score, covariate balancing scoring rule), and two weighting methods with balancing constraints but without using the propensity scores (entropy balancing and kernel balancing). We observed that correct specification of the propensity score model remains important even when the constraints effectively balance the covariates. We also observed evidence suggesting that, with similar covariate balance constraints, the use of a propensity score model improves the estimation performance when the dimension of covariates is large. These findings suggest that it is important to develop flexible data-driven propensity score models that satisfy covariate balancing conditions

Estimating Causal Effect of Restricted Mean Survival Time Difference based on Matched Design in Observational Survival Data

♦*Zihan Lin, Andy Ni and Bo Lu*
OSU
lu.232@osu.edu

Time to event outcome is commonly encountered in health and epidemiology research. Multiple papers have discussed the inadequacy of using hazard ratio as a causal effect measure due to its non-collapsibility and the time-varying nature. We adopt the restricted mean survival time (RMST) difference as a causal effect measure, since it essentially measures the mean discrepancy and has simple interpretation as the difference of areas under survival curves. To remove observed confounding, matching is used to pair similar treated and untreated subjects together, which is more robust to model misspecification. Simulation studies demonstrate that matching estimator has favorable performance as compared to other competing methods. Sensitivity analysis is also performed to assess the impact due to potential unmeasured confounding.

Using multiple imputation to classify potential outcomes subgroups

Yun Li
University of Pennsylvania
yun.li@pennmedicine.upenn.edu

With medical tests becoming increasingly available, concerns about over-testing, over-treatment and health care cost dramatically increase. Hence, it is important to understand the influence of testing on treatment selection in general practice. Most statistical methods focus on average effects of testing on treatment decisions. However, this may be ill-advised, particularly for patient subgroups that tend not to benefit from such tests. Furthermore, missing data are common, representing large and often unaddressed threats to the validity of most statistical methods. Finally, it is often desirable to conduct analyses that can be interpreted causally. Using the Rubin Causal Model framework, we propose to classify patients into four potential outcomes subgroups, defined by whether or not a patient's treatment selection is changed by the test result and by the direction of how the test result changes treatment selection. This subgroup classification naturally captures the differential influence of medical testing on treatment selections for different patients, which can suggest targets to improve the utilization of medical tests. We can then examine patient characteristics associated with patient potential outcomes subgroup memberships. We used multiple imputation methods to simultaneously impute the missing potential outcomes as well as regular missing values. This approach can also provide estimates of many traditional causal quantities of interest. We find that explicitly incorporating causal inference assumptions into the multiple imputation process can improve the precision for some causal estimates of interest. We also find that bias can occur when the potential outcomes conditional independence assumption is violated; sensitivity analyses are proposed to assess the impact of this violation. We applied the proposed methods to examine the influence of 21-gene assay, the most commonly used genomic test in the United States, on chemotherapy selection among breast cancer patients.

Regression-based causal inference with factorial experiments: estimands, model specifications, and design-based properties

Anqi Zhao[1] *and* ♦*Peng Ding*[2]
[1]National University of Singapore
[2]University of California Berkeley
pengdingpku@gmail.com

Factorial designs are widely used due to their ability to accommodate multiple factors simultaneously. The factor-based regression with main effects and some interactions is the dominant strategy for downstream data analysis, delivering point estimators and standard errors via one single regression. Justification of these convenient estimators from the design-based perspective requires quantifying their sampling properties under the assignment mechanism conditioning on the potential outcomes. To this end, we derive the sampling properties of the factor-based regression estimators from both saturated and unsaturated models, and demonstrate the appropriateness of the robust standard errors for the Wald-type inference. We then quantify the bias-variance trade-off between the saturated and unsaturated models from the design-based perspective, and establish a novel design-based Gauss–Markov theorem that ensures the latter's gain in efficiency when the nuisance effects omitted indeed do not exist. As a byproduct of the process, we unify the definitions of factorial effects in various literatures and propose a location-shift strategy for their direct estimation from factor-based regressions. Our theory and simulation suggest using factor-based inference for general factorial effects, preferably with parsimonious specifications in accordance with the prior knowledge of zero nuisance effects.

Session 21: Trial design and efficacy estimation of COVID-19 vaccine

Trial design and efficacy estimation of COVID-19 vaccine

Peter Gilbert
Fred Hutchinson Cancer Research Center
pgilbert@scharp.org

To be determined

Evaluating Vaccine Efficacy Against SARS-CoV-2 Infection

Danyu Lin
University of North Carolina
LIN@BIOS.UNC.EDU

Although interim results from several large placebo-controlled phase 3 trials demonstrated high vaccine efficacy (VE) against symptomatic COVID-19, it is unknown how effective the vaccines are in preventing people from becoming asymptomatically infected and potentially spreading the virus unwittingly. It is more difficult to evaluate VE against SARS-CoV-2 infection than against symptomatic COVID-19 because infection is not observed directly but rather is known to occur between two antibody or RT-PCR tests. Additional challenges arise as community transmission changes over time and as participants are vaccinated on different dates because of staggered enrollment of participants or crossover of placebo recipients to the vaccine arm before the end of the study. Here, we provide valid and efficient statistical methods for estimating potentially waning VE against SARS-CoV-2 infection with blood or nasal samples under time-varying community transmission, staggered enrollment, and blinded or unblinded crossover. We demonstrate the usefulness of the proposed methods through numerical studies mimicking the BNT162b2 phase 3 trial and the Prevent COVID U study. In addition, we assess how crossover and the frequency of diagnostic tests affect the precision of VE estimates.

Identifying COVID-19 vaccine correlates of protection from randomized trials

♦*David Benkeser*[1]*, Iván Díaz*[2] *and Jialu Ran*[3]
[1]Emory University
[2]Weill Cornell
[3]jialu.ran@emory.edu
benkeser@emory.edu

Combating the SARS-CoV2 pandemic will require the continued development of effective preventive vaccines. Regulatory agencies may open accelerated approval pathways for vaccines if a correlate of protection can be established. A correlate of protection is an immunological marker that is established as a strong predictor of the level of a vaccine's protection. A rich source of information for identifying such correlates are large-scale efficacy trials of COVID-19 vaccines, where immune responses are measured subject to a case-cohort sampling design. I will describe recent statistical advances for learning about vaccine correlates of protection using data from US government funded Phase 3 trials.

Vaccine Development During Pandemic: Innovative Design Transforming Development Paradigm

Satrajit Roychoudhury
Pfizer Inc
satrajit.roychoudhury@pfizer.com

Vaccines are complex biological products which are administered to healthy individuals. Safety is therefore paramount; vaccine development often entails large, time-consuming, and resource-intensive studies to detect rare safety issues and to establish vaccine efficacy. Before a vaccine is licensed and brought to the market, it undergoes a long and rigorous process of research, followed by many years of clinical testing. However, such framework requires modification for COVID-19 vaccine development due to high public health demand. This talk will present the operationally seamless development paradigm used to develop a mRNA vaccine for COVID-19. The design contains two parts. Phase 1 part was escalating dose levels in small cohorts in two age groups to identify a preferred candidate and dose level. It is followed by a phase 2/3 to evaluate safety, immunogenicity, and vaccine efficacy. A Bayesian group sequential designs was used for phase 2/3 part. This trial incorporates multiple interim analyses to assess early efficacy and futility of the vaccine. The Bayesian framework enabled us to obtain efficient designs using decision criteria based on the probability of benefit or harm. It also enabled us to incorporate information from previous studies on the treatment effect via the prior distributions. For COVID-19 vaccine trial, vaccine efficacy was based on achieving a sufficiently high Bayesian posterior probability. In addition, this trial has incorporated early stopping for futility based on Bayesian predictive probabilities. The talk will include key statistical aspects and regulatory challenges of the design. Publicly disclosed results will be summarized and discussed.

Session 22: Manifold Learning and Geometric Methods in Statistics

Principal Sub-manifolds and Classification on Manifolds

Zhigang Yao
National University of Singapore
zhigang.yao@nus.edu.sg

We will discuss the problem of finding principal components to the multivariate datasets, that lie on an embedded nonlinear Riemannian manifold within the higher-dimensional space. Our aim is to extend the geometric interpretation of PCA, while being able to capture the non-geodesic form of variation in the data. We introduce the concept of a principal sub-manifold, a manifold passing through the center of the data, and at any point of the manifold, it moves in the direction of the highest curvature in the space spanned by the eigenvectors of the local tangent space PCA. We show the principal sub-manifold yields the usual principal components in Euclidean space. We illustrate how to find, use and interpret the principal sub-manifold, with which a classification boundary can be defined for data sets on manifolds.

Intrinsic and extrinsic deep learning on manifolds

♦*Lizhen Lin*[1]*, Yihao Fang*[1]*, Ilsang Ohn*[1] *and Bayan Saparbayeva*[2]
[1]The University of Notre Dame
[2]The University of Rochester
lizhen.lin@nd.edu

In this talk, we will discuss both intrinsic and extrinsic deep neural network (DNN) models on the manifolds. An intrinsic DNN employs a Riemannian structure of the manifold while an extrinsic DNN relies on embedding a manifold onto a high-dimensional Euclidean space. The excessive risk of the DNN estimators will be derived and extensive numerical studies have been carried out to demonstrate the utilities of the models and illustrate the role of the geometry in developing the DNN models.

Gaussian process subspace regression: How to do PCA without a data sample?

♦*Ruda Zhang, Simon Mak and David Dunson*
Duke University
ruda.zhang@duke.edu

Subspace-valued functions arise in parametric reduced order modeling (PROM), where each parameter point is associated with a subspace, which is used for Petrov-Galerkin projections of large system matrices. Previous efforts to approximate such functions use interpolations on manifolds, which can be inaccurate and slow. We propose a matrix-valued Gaussian process (GP) to estimate subspace-valued functions, which we call GPS. This method is extrinsic and intrinsic at the same time: with multivariate Gaussian distributions on the Euclidean space, we induce a joint probability model on the Grassmann manifold, the set of fixed-dimensional subspaces. GPS adopts a simple yet general correlation structure,

and a principled approach for model selection. Its predictive distribution admits an analytical form and efficient methods for inference and sampling. For PROM problems, it gives a probabilistic prediction at a new parameter point that retains the accuracy of local reduced models, at a computational complexity that does not depend on system dimension, and thus is suitable for online computation. We give four numerical examples to compare our method versus subspace interpolation, as well as two methods that interpolate local reduced models. Overall, GPS is the most data efficient, more computationally efficient than subspace interpolation, and gives smooth predictions with uncertainty quantification. We also present an extension to approximate mappings that output positive semi-definite matrices, which is useful for conducting approximate principal component analysis (PCA) using PCA data instead of a data sample. The associated paper and an R package is available at: https://github.com/rudazhang/gpsr

Amplitude mean of trajectories on S2

Zhengwu Zhang
UNC Chapel Hill
zhengwu_zhang@unc.edu

The problems of analysis and modeling of spherical trajectories, that is, continuous longitudinal data on S2, are important in several disciplines. These problems are challenging for two reasons: (1) nonlinear geometry of S2 and (2) the presence of phase variability in given data. In this talk I will introduce our recent developments on handling these challenges. More specifically, we develop a geometric framework for separating phase variability from given trajectories, leaving only the shape or the amplitude variability. The key idea is to represent each trajectory with a pair of variables, a starting point, and a transported square-root velocity curve (TSRVC), a curve in the tangent (vector) space at the starting point. The space of all such curves forms a vector bundle and the L2 norm, along with the standard Riemannian metric on S2, provides a natural, warping-invariant metric on this vector bundle. This leads to an efficient algorithm for registration of trajectories, that is, phase-amplitude separation, and computational tools, such as clustering, sample means, and principal component analysis (PCA) of the two components separately. It also helps derive simple statistical models of phase-amplitude components of spherical trajectories. The work is demonstrated using two datasets: a set of bird-migration trajectories and a set of hurricane paths in the Atlantic ocean.

Session 23: Modern streaming Data Analysis: detection and identification

Industrial Analytics Research Facing Digital Transformation

Fugee Tsung
HKUST
season@ust.hk

This talk will present and discuss the challenges and opportunities that data science and analytics face in the era of digital transformation and the roles we play to drive such transformation. In particular, there is a big opportunity for industrial and business analytics, under the digital transformation paradigm, in order to further explore ways of creating value from data and big data. In addition, this talk will update the recent progress in our Quality and Data Analytics Lab on change detection in heterogeneous data streams.

An EWMA chart for High dimensional Process with Multi-class Out-of-Control Information via Random Forest Learning

Dongdong Xiang
East China Normal University
ddxiang@sfs.ecnu.edu.cn

Modern manufacturing and quality monitoring involve multi-class out-of-control(OOC) information from the training sample. It is essential to use such information during online monitoring data streams from complex processes. In this paper, a monitoring framework is designed by combing the random forest technique with the exponentially weighted moving average method for monitoring complex processes with multi-class out-of-control(OOC)information. To be specific, a process surveillance technique in the form of a control chart is proposed based on the probability that the online data being classified as an in-control (IC) sample, and the control chart triggers an alarm when the probability is lower than the control limit. Our numerical findings based on the Monte-Carlo simulation show that the proposed control chart performs more effectively than its competitors under various distributions and data types, especially for high dimensional cases when multi-class OOC information is known in advance. Moreover, the proposed method is illustrated with an application using the data related to the hard disk manufacturing processes.

Adaptive Process Monitoring Using Covariate Information

Kai Yang[1] *and* ♦*Peihua Qiu*[2]
[1]Graduate Assistant
[2]Professor and Founding Chair
pqiu@ufl.edu

Statistical process control (SPC) 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/performance variables are often affected in a complex way by many covariates, such as material, labor, weather conditions, social/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/performance variables and the covariates, shifts in the quality/performance 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/performance variables. Thus, it is challenging to properly use covariate information for process monitoring in a general setting. This paper suggests 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/performance variables, but it can react to a future shift in the quality/performance 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.

Item Pool Quality Control in Educational Testing via Compound Sequential Change Detection

Yunxiao Chen[1]*, Yi-Hsuan Lee*[2] *and* ♦*Xiaoou Li*[3]
[1]London School of Economics and Political Science
[2]Educational Testing Service
[3]University of Minnesota
lixx1766@umn.edu

In this talk, we introduce a compound sequential change detection framework and discuss its application in the monitoring of educational test item pools. This framework consists of (1) a multistream Bayesian change point model describing sequential changes in items, (2) a compound risk function quantifying the risk in sequential decisions, and (3) sequential decision rules that control the compound risk. Throughout the sequential decision process, the proposed decision rule balances the trade-off between two sources of errors, the false detection of pre-change items and the nondetection of post-change items. An item-specific monitoring statistic is proposed based on an item response theory model that eliminates the confounding from the examinee population which changes over time. Sequential decision rules and their theoretical properties are developed under two settings: the oracle setting where the Bayesian change point model is completely known and a more realistic setting where some parameters of the model are unknown.

Session 24: Robust methods for feature selection in high-dimensional problems

The Bayesian Regularized Quantile Varying Coefficient Model

♦*Cen Wu*[1]*, Fei Zhou*[1] *and Jie Ren*[2]
[1]Department of Statistics, Kansas State University
[2]Department of Biostatistics and Health Data Sciences, Indiana University School of Medicine
wucen@ksu.edu

In gene-environment interaction studies, the quantile varying coefficient (VC) models have played an important role in modelling nonlinear G-E interactions while being robust to heavy-tailed distributions and outliers in disease phenotypes. To date, variable selection for high dimensional quantile VC models have been extensively examined in the frequentist framework. In this talk, we present a study on the topic from a Bayesian perspective. Through the hierarchical model formulation, we have developed an efficient Gibbs sampler for the corresponding MCMC algorithm. In simulation, the merit of the proposed method has been demonstrated by comparisons with benchmark methods. In a case study of nonlinear G-E interactions, the Bayesian sparse quantile VC model yields promising identification and prediction results.

Recent developments in robust estimation and variable selection procedures

Irène Gijbels Gijbels
KU Leuven, Belgium
irene.gijbels@kuleuven.be

In this talk we present some recent ideas about a new class of robust M-estimators for performing simultaneously estimation and variable selection in high-dimensional regression models. We present the key ingredients of the procedure, involving penalization techniques. We provide some review on the use of nonconvex penalties and of nonconvex losses in a robust estimation context.

Robust Regression With Covariate Filtering: Heavy Tails and Adversarial Contamination

Ankit Pensia[1]*, Varun Jog*[2] *and* ♦*Po-Ling Loh*[2]
[1]UW-Madison
[2]University of Cambridge
pll28@cam.ac.uk

We study the problem of linear regression where both covariates and responses are potentially (i) heavy-tailed and (ii) adversarially contaminated. Several computationally efficient estimators have been proposed for the simpler setting where the covariates are sub-Gaussian and uncontaminated; however, these estimators may fail when the covariates are either heavy-tailed or contain outliers. In this work, we show how to modify the Huber regression, least trimmed squares, and least absolute deviation estimators to obtain estimators which are simultaneously computationally and statistically efficient in the stronger contamination model. Our approach is quite simple, and consists of applying a filtering algorithm to the covariates, and then applying the classical robust regression estimators to the remaining data. We show that the Huber regression estimator achieves near-optimal error rates in this setting, whereas the least trimmed squares and least absolute deviation estimators can be made to achieve near-optimal error after applying a postprocessing step.

Simultaneous Feature Selection and Outlier Detection with Optimality Guarantees

♦*Luca Insolia*[1]*, Ana Kenney*[2]*, Francesca Chiaromonte*[2] *and Giovanni Felici*[3]
[1]Sant'Anna School of Advanced Studies
[2]Pennsylvania State University
[3]National Research Council of Italy
Luca.insolia@sns.it

Sparse estimation in the presence of outliers has received considerable attention in the last decade. We contribute by considering high-dimensional regression models contaminated by multiple mean-shift outliers affecting both the response and the design matrix. We develop a general framework and use mixed-integer programming techniques to simultaneously perform feature selection and outlier detection with provably optimal guarantees. We prove theoretical properties for our approach, i.e., a necessary and sufficient condition for the robustly strong oracle property, where the number of features can increase exponentially with the sample size; the optimal estimation of parameters; and the breakdown point of the resulting estimates. Notably, our proposal requires weaker assumptions than prior methods in the literature and, unlike such methods, it allows the sparsity level and/or the amount of contamination to grow with the number of predictors and/or the sample size. Moreover, we provide computationally efficient procedures to tune integer constraints and warm-start the solution algorithm, and, through simulations, show the superior performance of our proposal with respect to existing heuristic methods. Finally, the method is deployed to elicit the role of microbiome in childhood obesity.

Session 25: Nonparametric Learning: New Directions and Innovations

Multiple domain and multiple kernel outcome-weighted learning for estimating individualized treatment regimes

Shanghong Xie[1]*, Thaddeus Tarpey*[2]*, Eva Petkova*[2] *and* ♦*Todd Ogden*[1]
[1]Columbia University
[2]NYU
to166@columbia.edu

Individualized treatment rules (ITRs) recommend treatment tailored specifically for each patient's own characteristics. It can be challenging to estimate optimal ITRs in the context of a large number of features, especially from multiple data domains (e.g., demographics, clinical measurements, neuroimaging modalities). Incorporating prior domain knowledge and using multiple similarity

measures to capture the potential complex relationship between features and treatment can potentially improve the accuracy of assigning treatment. Outcome weighted learning (OWL) methods that are based on support vector machine using a predetermined single kernel function have previously been developed to estimate optimal ITRs. We propose a new approach to estimate optimal ITRs by exploiting multiple kernel functions to describe the similarity of features between subjects both within and across data domains under the OWL framework, rather than preselecting a single kernel function to be used for all features. Our method takes account of the heterogeneity of each data domain and combines multiple data domains optimally. Also, it can estimate optimal ITRs and identify important data domains which can suggest priorities on the collection of data, potentially reducing cost without sacrificing accuracy. Relative advantages of several approaches are demonstrated by simulation studies and an application to a randomized clinical trial for major depressive disorder that collected features from multiple data domains.

Cross-sectional correlation tests with functional data and its application to inter- subject correlation analysis

Hongnan Wang and ♦*Ping-Shou Zhong*
University of Illinois at Chicago
pszhong@uic.edu

A focus of inter-subject correlation (ISC) analysis is to identify brain regions that respond similarly or synchronize to the same stimuli among a group of individuals by quantifying the inter-subject correlations. Functional MRI data are ideal for evaluating inter-subject correlation with continuous stimuli. We develop two nonparametric procedures to test the existence of cross-sectional correlation among functional curves. Both individual and temporal heterogeneity are allowed. We study the asymptotic distributions of the proposed test statistics under the null and local alternatives. Our empirical studies demonstrate that the proposed test procedure performs better than the commonly used methods in ISC studies, the adjusted Lagrange multiplier test, Pesaran's cross-sectional dependence (CD) test, and the adjusted Pesaran's CD test.

Sparse Modeling of Functional Linear Regression via Fused Lasso with Application to Genotype-by-environment Interaction Studies

♦*Shan Yu*[1]*, Aaron Kusmec*[2]*, Li Wang*[3] *and Dan Nettleton*[3]
[1]University of Virginia
[2]George Mason University
[3]Iowa State University
sy5jx@virginia.edu

We propose a sparse multi-group functional linear regression model to simultaneously estimate multiple coefficient functions and identify groups, such that coefficient functions are identical within groups and distinct across groups. By borrowing information from relevant subgroups of subjects, our method enhances estimation efficiency while preserving heterogeneity in model parameters and coefficient functions. We use an adaptive fused lasso penalty to shrink coefficient estimates to a common value within each group. We also establish theoretical properties of the proposed estimators. To enhance computation efficiency and incorporate neighborhood information, we propose to use graph-constrained adaptive lasso with a highly efficient algorithm. Two Monte Carlo simulation studies have been conducted to study the finite-sample performance of the proposed method. The proposed method is applied to sorghum flowering-time data and hybrid maize grain yields from the Genomes to Fields consortium.

Partial Separability and Functional Graphical Models

Javier Zapata[1]*, Sang-Yun Oh*[1] *and* ♦*Alexander Petersen*[2]
[1]University of California Santa Barbara
[2]Brigham Young University
petersen@stat.byu.edu

The covariance structure of multivariate functional data can be highly complex, especially if the multivariate dimension is large, making extension of statistical methods for standard multivariate data to the functional data setting quite challenging. For example, Gaussian graphical models have recently been extended to the setting of multivariate functional data by applying multivariate methods to the coefficients of truncated basis expansions. However, a key difficulty compared to multivariate data is that the covariance operator is compact, and thus not invertible. The methodology in this paper addresses the general problem of covariance modeling for multivariate functional data, and functional Gaussian graphical models in particular. As a first step, a new notion of separability for multivariate functional data is proposed, termed partial separability, leading to a novel Karhunen-Loève-type expansion for such data. Next, the partial separability structure is shown to be particularly useful in order to provide a well-defined Gaussian graphical model that can be identified with a sequence of finite-dimensional graphical models, each of fixed dimension. This motivates a simple and efficient estimation procedure through application of the joint graphical lasso. Empirical performance of the method for graphical model estimation is assessed through simulation and analysis of functional brain connectivity during a motor task.

Session 26: Challenges in the analysis of complex structured data

Cox regression with dynamic markers measured at covariate-dependent visit times

♦*Richard Cook and Bingfeng Xie*
University of Waterloo
rjcook@uwaterloo.ca

Cox regression is routinely used to assess the association between internal time-dependent covariates (markers) and failure times. The markers are typically only measured at clinic visits when careful examination can take place, or when blood samples are drawn for testing, so they are strictly under intermittent observation. Examples include measurement of blood glucose control in diabetics , blood pressure in hypertension trials, or markers of bone destruction in patients with skeletal metastases. We consider the asymptotic bias of estimators of the marker effects on the hazard for failure when the marker values are carried forward until the next clinic visit. We also consider the setting where the marker values, often reflecting the severity of disease activity, affect the intensity for clinic visits. A joint model is then proposed for the marker, visit, and failure process to mitigate this bias. Simulations are reported on to investigate the finite sample performance of naive and valid estimators and a clinical illustration is provide.

Semiparametric Regression Model for Ordinal Response

Ao Yuan
Georgetown University
ay312@georgetown.edu

Ordinal data with covariates is omnipresent in practice. Methods have relied on subjectively specified models, and it is known that when the specified model deviates from the true one, the resulting estimates are not reliable, especially for prediction. Although existing semiparametric ordinal models have improved upon parametric

models by trying to account for potential nonlinear effects of the covariates, these models are still not flexible enough. We propose a broadly applicable robust semiparametric ordinal regression model, in which the response is a nonparametric monotone increasing function of the parametrically specified covariates for robustness.

Risk Projection for Time-to-Event Leveraging Summary Statistics With Source Individual-level Data

Jiayin Zheng, Yingye Zheng and ♦*Li Hsu*
Fred Hutchinson Cancer Research Center
lih@fredhutch.org

Predicting risks of chronic diseases has become increasingly important in clinical practice. When a prediction model is developed in a cohort, there is a great interest to apply the model to other cohorts. Due to potential discrepancy in baseline disease incidences between different cohorts and shifts in patient composition, the risk predicted by the model built in the source cohort often under- or over-estimates the risk in a new cohort. We develop a weighted estimating equation approach to re-calibrating the projected risk for the targeted population. The recalibration leverages the knowledge about disease incidence rates and some summary information of risk factors in the target population. Through extensive simulation we demonstrate that the proposed estimators are robust, even if the risk factor distributions differ between the source and target populations, and gain efficiency if they are the same if the information from the target is precise. Finally, we illustrate the method with recalibration of a colorectal cancer prediction model.

A multiple imputation method for nonlinear mixed effects models with missing data

Lang Wu
University of British Columbia
lang@stat.ubc.ca

Multiple imputation methods are widely used in practice for missing data. An important consideration for a multiple imputation method is the choice of an imputation model which generates the imputations for each missing value, especially when the missing rate is not low. Mixed effects models are commonly used for modelling longitudinal data which exhibit large between-individual variations. In this case, a good imputation model should generate imputations at the individual level to incorporate the large between-individual variations. In this talk, we propose a multiple imputation method for nonlinear mixed effects models with missing responses. We consider an iterative linearization method where the imputations are generated based on a working linear mixed effects model. We also discuss order-restricted hypothesis testing in nonlinear mixed effects models with missing data based on the proposed multiple imputation method.

Session 27: Recent development on the analysis of single-cell RNA-seq data

Flexible experimental designs for valid single-cell RNA-sequencing experiments allowing batch effects correction

♦*Fangda Song*[1] *and Yingying Wei*[2]
[1]The Chinese University of Hong Kong, Shenzhen
[2]The Chinese University of Hong Kong
songfd2016@link.cuhk.edu.hk

Despite their widespread applications, single-cell RNA-sequencing (scRNA-seq) experiments are still plagued by batch effects and dropout events. Although the completely randomized experimental design has frequently been advocated to control for batch effects, it is rarely implemented in real applications due to time and budget constraints. Here, we mathematically prove that under two more flexible and realistic experimental designs-the reference panel and the chain-type designs-true biological variability can also be separated from batch effects. We develop Batch effects correction with Unknown Subtypes for scRNA-seq data (BUSseq), which is an interpretable Bayesian hierarchical model that closely follows the data-generating mechanism of scRNA-seq experiments. BUSseq can simultaneously correct batch effects, cluster cell types, impute missing data caused by dropout events, and detect differentially expressed genes without requiring a preliminary normalization step. We demonstrate that BUSseq outperforms existing methods with simulated and real data.

Model-Based Trajectory Inference for Single-Cell RNA Sequencing Using Deep Learning with a Mixture Prior

Jinhong Du, Ming Gao and ♦*Jingshu Wang*
University of Chicago
jingshuw@uchicago.edu

Trajectory inference methods analyze thousands of cells from single-cell sequencing technologies and computationally infer their developmental trajectories. Though many tools have been developed for trajectory inference, most of them lack a coherent statistical model and reliable uncertainty quantification. In this talk, I'll present VITAE, a statistical method that combines a latent hierarchical mixture model with variational autoencoders to infer trajectories from posterior approximations. VITAE is computationally scalable and can adjust for confounding covariates to integrate multiple datasets. We show that VITAE outperforms other state-of-the-art trajectory inference methods on both real and synthetic data under various trajectory topologies. We also apply VITAE to jointly analyze two single-cell RNA sequencing datasets on the mouse neocortex. Our results suggest that VITAE can successfully uncover a shared developmental trajectory of the projection neurons and reliably order cells from both datasets along the inferred trajectory.

Estimating Heterogeneous Gene Regulatory Networks from Zero-Inflated Single-Cell Expression Data

♦*Xiangyu Luo and Qiuyu Wu*
Renmin University of China
xiangyuluo@ruc.edu.cn

Inferring gene regulatory networks can elucidate how genes work cooperatively. The gene-gene collaboration information is often learned by Gaussian graphical models (GGM) that aim to identify whether the expression levels of any pair of genes are dependent given other genes' expression values. One basic assumption that guarantees the validity of GGM is data normality, and this often holds for bulk-level expression data which aggregate biological signals from a collection of cells. However, fine-grained cell-level expression profiles collected in single-cell RNA-sequencing (scRNA-seq) reveal non-normality features—cellular heterogeneity and zero-inflation. We propose a Bayesian latent mixture GGM to jointly estimate multiple gene regulatory networks accounting for the zero-inflation and unknown heterogeneity of single-cell expression data. The proposed approach outperforms competing methods on synthetic data in terms of network structure and precision matrix estimation accuracy and provides biological insights when applied to two real-world scRNA-seq datasets.

Non-Parametric Modeling Enables Scalable and Robust Detection of Spatial Expression Patterns for Large Spatial Transcriptomic Studies

Xiang Zhou

University of Michigan
xzhousph@umich.edu
Spatial transcriptomic studies are becoming increasingly common and increasingly large, posing important statistical and computational challenges for many analytic tasks. Here, we present SPARK-X, a non-parametric modeling framework for rapid and effective identification of genes with spatial expression patterns in large spatial transcriptomic studies. SPARK-X not only produces effective type I error control and high power but also brings orders of magnitude computational savings. We applied SPARK-X to analyze three large spatial transcriptomic data, one of which is only analyzable by SPARK-X. In these data, SPARK-X identified many spatially expressed genes not identifiable by existing approaches, revealing new biological insights.

Session 28: AI Boosting of pharmaceutical drug development and the emerging world of digital therapeutics

Subgroup analysis based on K-means for multivariate longitudinal data

♦*Gen Zhu*[1]*, Margaret Gamalo*[2] *and Chuanbo Zang*[2]
[1]UTHealth
[2]Pfizer
gen.zhu@uth.tmc.edu
Subgroup analysis is important to the clinical trial community to establish robustness of response of treatments and to inform its use. Typically, subgroup investigations are based on one response at one timepoint and baseline patient characteristics to identify meaningful biomarkers even if there is available data at multiple timepoints. In this study, we employed the K-means algorithm to cluster the patients into subgroups according to their longitudinal responses profile at all post-baseline timepoints . A simulation experiment showed that the method performed well on recovering the true subgroups. The method was then applied to the data from a clinical trial study. We found patients given the treatment were clustered into two groups, the enhanced benefit group and the moderate benefit group. We also investigated different machine learning methods to predict whether a patient can achieve enhanced benefit at later stages by using the baseline characteristics and early response.

Beyond step counting-Measuring Nighttime Scratch and Sleep with Wearable Devices

Nikhil Mahadevan, Yiorgos Christakis, ♦*Junrui Di, Jonathan Bruno, Yao Zhang, Carrie Northcott and Shyamal Pate*
Pfizer
Junrui.Di@pfizer.com
Patients with atopic dermatitis experience increased nocturnal pruritus which leads to scratching and sleep disturbances that significantly contribute to reductions in their quality of life. Objective measurements of nighttime scratching and sleep quantity can help assess the efficacy of an intervention. Wearable sensors can provide novel, objective measures of nighttime scratching and sleep; however, many current approaches were not designed for passive, unsupervised monitoring during daily life. In this work, we present the development and analytical validation of a machine learning-based system that sequentially processes epochs of sample-level accelerometer data from a wrist-worn device to provide continuous digital measures of night-time scratching and sleep quantity. This approach uses heuristic and machine learning algorithms in a hierarchical paradigm by first determining when the patient intends to sleep, then detecting sleep-wake states along with scratching episodes, and lastly deriving objective measures of both sleep and scratch. When these digital endpoints are monitored in conjunction with pharmacotherapies it may improve overall treatment of the condition, as well as increase our understanding of these key symptoms and impact of the pharmacotherapy on them.

Statistical Considerations for the Clinical Validations of AI/ML-based Digital Health Products

Feiming Chen
Food and Drug Administration
Feiming.Chen@fda.hhs.gov
Artificial Intelligence (AI) and Machine-Learning (ML) algorithms have been increasingly used in the medical products, such as Software as a Medical Device (SaMD). We consider the statistical aspects in evaluating such products, with a focus on medical diagnostic devices. We provide relevant academic references and advocate good practices in addressing various statistical challenges in the clinical validation of AI/ML-based medical devices.

Session 29: Real World Evidence Innovation using Causal Methodology and Application

Propensity Score Methods in for Multiple Treatment Comparisons for Evaluating Drug Safety and Effectiveness

Rongmei Zhang
Center for Drug Evaluation and Research
rongmei.zhang@fda.hhs.gov
The rapid advances in technology and vast increase in the amount of real world data create an unprecedented opportunity to use real world evidence to inform patient care, supplement current clinical trials, and improve post-marketing drug monitoring and evaluation. Propensity score methods have been widely used to reduce confounding and examine the causal effects of treatment in non-randomized studies. Most studies used propensity score methods to compare two treatment groups of interests, while methods and guidance for propensity score methods when examining multiple treatments are limited. This talk will discuss the challenges of multiple treatments comparisons with real word data, present our propensity score methods, and share our experiences using examples in post-market comparative effectiveness studies.

Comparison Study of Causal Methods for Average Treatment Effect Estimation Allowing Covariate Measurement Error

♦*Yi Huang*[1] *and Feng Zhou*[2]
[1]UMBC
[2]FDA
yihuang@umbc.edu
Propensity score approaches are widely used in real world applications to draw causal inference on average treatment effect (ATE) and to quantify real world evidence to support and help FDA on regulatory decision making. However, covariate measurement error is a well-known challenge using observational studies drawing causal inference due to the violation of un-confoundedness assumption. Ignoring measurement error and using naïve propensity scores lead to biased ATE estimates. Four causal methods are found that can control the influence of covariate measurement error without external data, and there is no existing literature comparing their numerical performances: Battistin and Chesher's bias correction method, MaCaffrey and Lockwood's inverse probability weighting method, and our latent propensity score methods estimated by EM algorithm and MCMC algorithm. Systematic simulation studies are conducted to compare their performances under rationales with respect to Gaussian vs. binary outcome, continuous vs. discrete un-

derlying true covariate, small vs. large treatment effect, and small vs. large measurement error. The results show that under Gaussian outcome, the bias correction method and the latent propensity score method using EM (expectation-maximization) algorithm perform best with small and large measurement error respectively; under binary outcome, the inverse probability weighting method and the latent propensity score method using MCMC algorithm perform best with small and large measurement error respectively. This is a joint work with former PhD student Zhou Feng.

Real World Data for Clinical Development: Clinical Eligibility Criteria, Hybrid Control Arms, and Challenges

Thomas Jemielita
Merck & Co., Inc.
thomas.jemielita@merck.com

Following the 20th-century cures act, there has been increased interest in using real-world data (RWD) to enhance or augment decision-making for clinical development. For example, RWD could be used to benchmark a single-arm study or augment the clinical control arm. While there are promising applications for RWD, there are considerable challenges such as confounding due to using non-randomized data. Importantly, clinical trial patients are enrolled based on pre-specified eligibility criteria, and in an ideal world, we would select RWD patients similarly. In practice, it is unlikely that all eligibility criteria are available in the RWD, and some eligibility criteria (ex: life expectancy) are not directly measurable in any data source. With many eligibility criteria and missing information, this can rapidly shrink the available sample size. Another promising avenue for RWD is "hybrid control arms." For example, a sponsor may randomize patients to the study drug and study control through a 2:1 randomization ratio and further augment with RWD control patients. Compared to a single-arm setting (ex: study treatment vs RWD control), this approach can directly leverage the clinical control to "adjust" the RWD control towards the target population of interest. Overall, this talk will discuss these topics in-depth, along with potential solutions and real-data examples.

Optimal Treatment Regimes: An Empirical Comparison of Methods and Applications

Zhen Li[1], ♦*Jie Chen*[2], *Eric Labor*[3], *Fang Liu*[4] *and Richard Baumgartner*[4]
[1]Amazon
[2]Overland Pharma
[3]Duke University
[4]Merck
jchen@overlandpharma.com

A treatment regime is a sequence of decision rules, one per decision point, that maps accumulated patient information to a recommended intervention. An optimal treatment regime maximizes expected cumulative utility if applied to select interventions in a population of interest. As a treatment regime seeks to improve the quality of healthcare by individualizing treatment, it can be viewed as an approach to formalizing precision medicine. Increased interest and investment in precision medicine have led to a surge of methodological research focusing on estimation and evaluation of optimal treatment regimes from observational and/or randomized studies. These methods are becoming commonplace in biomedical research, though guidance about how to choose among existing methods in practice has been somewhat limited. The purpose of this presentation is to describe some of the most commonly used methods for estimation of an optimal treatment regime, and to compare these estimators in a series of simulation experiments and application to real data. The results of these simulations along with the theoretical/methodological properties of these estimators are used to form recommendations for applied researchers.

Session 30: Recent development in Pediatric Clinical Trial Design

Causal Inference in Rare Diseases

Rima Izem
Novartis
rima.izem@novartis.com

Most rare diseases have a genetic cause and impact children. Sound study design and causal inference methods are essential to demonstrate the therapeutic efficacy, safety, and effectiveness of new therapies. In the rare diseases setting, several factors challenge the use of typical parallel control designs: the small patient population size, its genotypic and phenotypic diversity, and the complexity and incomplete understanding of the disorder's progression. Designs with repeated measures in longitudinal studies can increase study power and reduce heterogeneity. This paper reviews these designs and draws the parallel between some new and existing randomized studies in rare diseases and their less well-known controlled observational study designs. We show that self-controlled randomized crossover and N-of-1 designs have similar considerations as the observational case series and case-crossover designs. Also, randomized sequential designs have similar considerations to longitudinal cohort studies using sequential matching or weighting to control confounding. We discuss design and analysis considerations for valid causal inference and illustrate them with examples of analyses in multiple rare disorders, including urea cycle disorder.

The Potential of Master Protocols in Pediatric Clinical Trials

Kristine Broglio
AstraZeneca
kristine.broglio@astrazeneca.com

A master protocol is a clinical trial protocol designed to answer multiple questions under an overarching infrastructure. Three types of master protocol trials have been described including umbrella trials, platform trials, and basket trials. This talk will focus on platform trials, which are master protocol studies used to simultaneously investigate multiple different candidate therapies for the same patient population. Platform trials have been shown to provide statistical efficiencies including reducing total sample size, total trial duration, and the number of patients assigned to a control arm. This talk will review the current landscape of platform trials, describe their unique statistical and operational considerations, and highlight the potential benefits of this approach for pediatric drug development.

Integrating adult data into design of pediatric dose-finding studies

♦*Yimei Li*[1] *and Ying Yuan*[2]
[1]University of Pennsylvania
[2]University of Texas MD Anderson Cancer Center
LiY3@chop.edu

Pediatric phase I trials are usually carried out after the adult trial has started, but not completed yet. As the pediatric trial progresses, in light of the accrued interim data from the concurrent adult trial, the pediatric protocol often is amended to modify the original pediatric dose escalation design. This frequently is done in an ad hoc way, interrupting patient accrual and slowing down the trial. We develop a pediatric continuous reassessment method (PA-CRM) to streamline this process, providing a more efficient and rigorous method

to find the MTD for pediatric phase I trials. We use a discounted joint likelihood of the adult and pediatric data, with a discount parameter controlling information borrowing between pediatric and adult trials. According to the interim adult and pediatric data, the discount parameter is adaptively updated using the Bayesian model averaging method. We examine the PA-CRM through simulations, and compare it with the two alternative approaches, which ignore adult data completely or simply pool it together with the pediatric data. The results demonstrate that the PA-CRM has good operating characteristics and is robust to various assumptions.

Session 31: Recent development in machine learning and causal inference

Conformal Inference of Counterfactuals and Individual Treatment Effects

♦*Lihua Lei and Emmanuel Candès*
Stanford University
lihualei@stanford.edu

Evaluating treatment effect heterogeneity widely informs treatment decision making. At the moment, much emphasis is placed on the estimation of the conditional average treatment effect via flexible machine learning algorithms. While these methods enjoy some theoretical appeal in terms of consistency and convergence rates, they generally perform poorly in terms of uncertainty quantification. This is troubling since assessing risk is crucial for reliable decision-making in sensitive and uncertain environments. In this work, we propose a conformal inference-based approach that can produce reliable interval estimates for counterfactuals and individual treatment effects under the potential outcome framework. For completely randomized or stratified randomized experiments with perfect compliance, the intervals have guaranteed average coverage in finite samples regardless of the unknown data generating mechanism. For randomized experiments with ignorable compliance and general observational studies obeying the strong ignorability assumption, the intervals satisfy a doubly robust property which states the following: the average coverage is approximately controlled if either the propensity score or the conditional quantiles of potential outcomes can be estimated accurately. Numerical studies on both synthetic and real datasets empirically demonstrate that existing methods suffer from a significant coverage deficit even in simple models. In contrast, our methods achieve the desired coverage with reasonably short intervals.

Optimal and Safe Estimation for High-Dimensional Semi-Supervised Learning

Jiwei Zhao
University of Wisconsin Madison
jiwei.zhao@wisc.edu

We consider the estimation problem in high-dimensional semi-supervised learning. Our goal is to investigate when and how the unlabeled data can be exploited to improve the estimation of the regression parameters of linear model in light of the fact that such linear models may be misspecified in data analysis. We first establish the minimax lower bound for parameter estimation in the semi-supervised setting. We show that the supervised estimators using the labeled data only cannot attain this lower bound. When the conditional mean function is correctly specified, we propose an optimal semi-supervised estimator which attains the lower bound and therefore improves the rate of the supervised estimators. To alleviate the strong requirement for this optimal estimator, we further propose a safe semi-supervised estimator. We view it safe, because this estimator remains minimax optimal when the conditional mean function is correctly specified, and is always at least as good as the supervised estimators. Furthermore, we extend our idea to aggregate multiple semi-supervised estimators caused by different misspecifications of the conditional mean function. Extensive numerical simulations and a real data analysis are conducted to illustrate our theoretical results. This is a joint work with Siyi Deng, Yang Ning and Heping Zhang.

Doubly Robust Semiparametric Inference Using Regularized Calibrated Estimation with High-dimensional Data

Zhiqiang Tan
Rutgers University
ztan@stat.rutgers.edu

Consider semiparametric estimation where a doubly robust estimating function for a low-dimensional parameter is available, depending on two working models. With high-dimensional data, we develop regularized calibrated estimation as a general method for estimating the parameters in the two working models, such that valid Wald confidence intervals can be obtained for the parameter of interest under suitable sparsity conditions if either of the two working models is correctly specified. We propose a computationally tractable two-step algorithm and provide rigorous theoretical analysis which justifies sufficiently fast rates of convergence for the regularized calibrated estimators in spite of sequential construction and establishes a desired asymptotic expansion for the doubly robust estimator. As concrete examples, we discuss applications to partially linear, log-linear, and logistic models and estimation of average treatment effects. Numerical studies in the former three examples demonstrate superior performance of our method, compared with debiased Lasso.

Profile Matching for the Generalization and Personalization of Causal Inferences

Eric Cohn and ♦*Jose Zubizarreta*
Harvard University
zubizarreta@hcp.med.harvard.edu

We introduce profile matching, a multivariate matching method for randomized experiments and observational studies that finds the largest possible self-weighted samples across multiple treatment groups that are balanced relative to a covariate profile. This covariate profile can represent a specific population or a target individual, facilitating the tasks of generalization and personalization of causal inferences. For generalization, because the profile often amounts to summary statistics for a target population, profile matching does not require accessing individual-level data, which may be unavailable for confidentiality reasons. For personalization, the profile can characterize a single patient. Profile matching achieves covariate balance by construction, but unlike existing approaches to matching, it does not require specifying a matching ratio, as this is implicitly optimized for the data. The method can also be used for the selection of units for study follow-up, and it readily applies to multivalued treatments with many treatment categories. We evaluate the performance of profile matching in a simulation study of generalization of a randomized trial to a target population. We further illustrate this method in an exploratory observational study of the relationship between opioid use treatment and mental health outcomes. We analyze these relationships for three covariate profiles representing: (i) sexual minorities, (ii) the Appalachian United States, and (iii) a hypothetical vulnerable patient. We provide R code with step-by-step explanations to implement the methods in the paper in the Supple-

mentary Materials.

Session 32: Prediction and inference for neural-network-based methods and their applications to genetics

A New Kernel Neural Network Method for Genetic Data Analysis

♦*Qing Lu*[1]*, Xiaoxi Shen*[1] *and Xiaoran Tong*[2]
[1]University of Florida
[2]NIH
lucienq@ufl.edu

Artificial intelligence (AI) is a thriving research field with many successful applications in areas such as computer vision and speech recognition. Neural-network-based methods (e.g., deep learning) play a central role in modern AI technology. While neural-network-based methods also hold great promise for genetic research, the high-dimensionality of genetic data, the massive amounts of study samples, and complex relationships between genetic variants and disease outcomes bring tremendous analytic and computational challenges. To address these challenges, we propose a kernel neural network (KNN) method. KNN inherits features from both linear mixed models (LMM) and classical neural networks and is designed for high-dimensional genetic data analysis. Unlike the classic neural network, KNN summarizes a large number of genetic variants into kernel matrices and uses the kernel matrices as input matrices. Based on the kernel matrices, KNN builds a feedforward neural network to model the complex relationship between genetic variants and a disease outcome. Minimum norm quadratic unbiased estimation and batch training are implemented in KNN to accelerate the computation, making KNN applicable to massive datasets with millions of samples. Through simulations, we demonstrate the advantages of KNN over LMM in terms of prediction accuracy and computational efficiency. We also apply KNN to the large-scale UK Biobank dataset, evaluating the role of a large number of genetic variants on multiple complex diseases.

Expectile Neural Networks for Genetic DataAnalysis of Complex Diseases

♦*Jinghang Lin*[1]*, Xiaoran Tong*[2]*, Chenxi Li*[1] *and Qing Lu*[3]
[1]Michigan State University
[2]National Institutes of Health
[3]University of Florida
linjingh@msu.edu

The genetic etiologies of common diseases are highly complex and heterogeneous. Classic methods, such as linearregression, have successfully identified numerous variants associated with complex diseases. Nonetheless, for most diseases, theidentified variants only account for a small proportion of heritability. Challenges remain to discover additional variants contributing tocomplex diseases. Expectile regression is a generalization of linear regression and provides complete information on the conditionaldistribution of a phenotype of interest. While expectile regression has many nice properties, it has been rarely used in geneticresearch. In this paper, we develop an expectile neural network (ENN) method for genetic data analyses of complex diseases. Similarto expectile regression, ENN provides a comprehensive view of relationships between genetic variants and disease phenotypes andcan be used to discover variants predisposing to sub-populations. We further integrate the idea of neural networks into ENN, making itcapable of capturing non-linear and non-additive genetic effects (e.g., gene-gene interactions). Through simulations, we showed thatthe proposed method outperformed an existing expectile regression when there exist complex genotype-phenotype relationships. We also applied the proposed method to the data from the Study of Addiction: Genetics and Environment(SAGE) investigating the relationships of candidate genes with smoking quantity.

A Goodness-of-Fit Test Based on Neural Networks

Xiaoxi Shen and ♦*Chang Jiang*
University of Florida
jiangc@ufl.edu

Neural networks have become one of the most popularly used methods in machine learning and artificial intelligence. Due to the universal approximation theorem (Hornik, Stinchcombe and White,1989), a neural network with one hidden layer can approximate any continuous function on compact support as long as the number of hidden units is sufficiently large. Statistically, a neural network can be classified into a nonlinear regression framework. However, if we consider it parametrically, due to the unidentifiability of the parameters, it is difficult to derive its asymptotic properties. Instead, we consider the estimation problem in a nonparametric regression framework and use the results from sieve estimation to establish the consistency, the rates of convergence and the asymptotic normality of the neural network estimators. We also illustrate the validity of the theories via simulations.

Multimodal Functional Deep Learning for Multi-omics Data

♦*Yuan Zhou*[1]*, Shan Zhang*[2]*, Pei Geng*[3] *and Qing Lu*[1]
[1]University of Florida
[2]Michigan State University
[3]Illinois State University
yuanzhou1@ufl.edu

With rapidly evolving high-throughput technologies and ever-decreasing costs, it becomes feasible to collect diverse types of omics data in large-scale studies. While the multi-omics data generated from these studies hold great promise for innovative insights on biology mechanisms of human disease, the high-dimensionality of omics data and the complexity between various levels of omics data and disease phenotypes bring tremendous analytic challenges. To address these challenges and to facilitate ongoing multi-omics analysis, we propose a Multimodal Functional Deep Learning (MFDL) method for high-dimensional multi-omics data analysis. MFDL model the complex relationships between genetic variants and disease phenotypes through the hierarchical structure of deep neural networks and handle high-dimensional omics data by using the functional data analysis technique. Moreover, MFDL utilizes the structure of the multimodal model to model interactions between multi-omics data. Through simulation studies and real data applications, we demonstrate the advantages of MFNN in terms of prediction accuracy as well as being robust to the high dimensionality and noise of the data.

Session 33: The fad and fade in statistics for biopharmaceutical research: editor?s pick from top biopharmaceutical statistics journals

BIOSTATISTICS RESEARCH TRENDS IN BIOMETRICS, THE FLAGSHIP JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY, IN PRE-, PERI-, AND POST-PANDEMIC TIMES

Geert Molenberghs
UHasselt & KU Leuven
geert.molenberghs@uhasselt.be

The journal Biometrics was founded in 1945, right after World War II, thus slightly predating the foundation of the International Biometric Society, of which it became the flagship journal. It subscribes to the mission of the IBS, which is being "An international society devoted to the development and application of statistical and mathematical theory and methods in the biosciences." The founding Editor was Gertrude M. Cox. At its foundation, a majority of work was devoted to non-medical biostatistics, such as in agriculture, forestry, and biology. Medical statistics quickly found its place in the journal, as well as research in observational studies, spurred by the growth of epidemiology. Ever since, the journal has shown its flexibility in embracing important new developments and giving them a proper place, without neglecting existing themes. Such trends include statistical genetics and bioinformatics around the turn of the millennium, and data science, to name but a few. The journal has had its share of confusion over its name - is it now biometry, biometrics, biostatistics, Related to this, the "other biometrics" or security biometrics emerged, a large part of which falls outside of the journal's remit. Over the most recent one and half year, the journal has welcomed a large number of COVID-19 related manuscripts, and even an increase in submissions as a whole, in contrast to many journals in other branches of science. Given the large number of journals in statistics, data science, computational biology, and related areas, there is of course overlap between Biometrics' remit and that of other journal. Rather than being a problem, this stimulates healthy competition for quality.

Not-so-popular stories - statistical methods that are not in the mainstream of biopharmaceutical statistics literature

Margaret Gamalo
Pfizer
Margaret.Gamalo@pfizer.com

Statistical literature is still influenced by what is "hot" and who is influential. In this talk, I will talk about researches that are off the beaten path and introduce problems that many statisticians neglect or are not aware of. I will also provide comments and how they can be improved as a future research direction.

Statistics in Medicine: what's new and what's not new?

Robert Platt
McGill University
robert.platt@mcgill.ca

Statistics in Medicine aims to influence practice in medicine and its associated sciences through the publication of papers on statistical and other quantitative methods. Over the years, we have seen "hot" areas come and go, some with major influence on the practice of statistics in medicine, and others less so. I will reflect on some of these areas, and point out some areas that changed things, and others that turned out to be hype. I will then talk about where the journal and the field is going, and point out some areas where we should invest in the future.

Out of its infancy: Statistics in Biopharmaceutical Research

♦*Frank Bretz*[1] *and Toshimitsu Hamasaki*[2]
[1]Novartis Pharma AG
[2]George Washington University
frank.bretz@novartis.com

Statistics in Biopharmaceutical Research (SBR) is a relatively new journal. The idea for SBR was originally proposed by Professor Bradley Efron during his term as the President of the American Statistical Association (ASA), and approved by the ASA Board of Directors in March 2006. SBR published its first issue in January 2009, and as of December 2020, 48 issues with approximately 500 contributions directed to researchers and applied statisticians from academia, government, and industry supporting the growing disciplines in the biopharmaceutical sciences. In this presentation, we review briefly the history of SBR. We then discuss emerging topics and how SBR embraces the changing landscape. Finally, we provide advice for potential contributors who would like to submit manuscripts to SBR.

Session 34: Recent Development in Multi-Block Data Integration

Simultaneous Clustering and Estimation of Networks via Sparse Tensor Decomposition

♦*Gen Li*[1] *and Miaoyan Wang*[2]
[1]University of Michigan, Ann Arbor
[2]University of Wisconsin, Madison
ligen@umich.edu

The standard Gaussian graphical models have been widely used to investigate the dependency structure among variables in a single population. As it is increasingly common to collect data from multiple heterogeneous populations, multi-layered networks become prevalent. In this paper, we consider the simultaneous clustering and estimation of multiple graphical models. We build upon the Gaussian graphical models and utilize a sparse tensor decomposition approach to simultaneously cluster populations and estimate the underlying network structures among variables in each population. A penalized likelihood method is used where we devise an alternating direction method of multipliers algorithm to estimate model parameters. We demonstrate the efficacy of the proposed method with comprehensive simulation studies. The application to the GTEx multi-tissue gene expression data provides important insights into tissue clustering and gene co-expression patterns in different tissues.

Semi-Supervised Statistical Inference for High-Dimensional Linear Regression with Blockwise Missing Data

♦*Fei Xue*[1]*, Rong Ma*[2] *and Hongzhe Li*[2]
[1]Purdue University
[2]University of Pennsylvania
fei.xue@pennmedicine.upenn.edu

Blockwise missing data occurs frequently when we integrate multisource or multimodality data where different sources or modalities contain complementary information. In this paper, we consider a high-dimensional linear regression model with blockwise missing covariates and a partially observed response variable. Under this semi-supervised framework, we propose a computationally efficient estimator for the regression coefficient vector based on carefully constructed unbiased estimating equations and a multiple blockwise imputation procedure, and obtain its rates of convergence. Furthermore, building upon an innovative semi-supervised projected estimating equation technique that intrinsically achieves bias-correction of the initial estimator, we propose nearly unbiased estimators for the individual regression coefficients that are asymptotically normally distributed under mild conditions. By carefully analyzing these debiased estimators, asymptotically valid confidence intervals and statistical tests about each regression coefficient are constructed. Numerical studies and application analysis of the Alzheimer's Disease Neuroimaging Initiative data show that the proposed method performs better and benefits more from unsupervised samples than existing methods.

Joint Matrix Decomposition Regression for Multi-omics Data Analysis

♦*Yue Wang*[1]*, Jing Ma*[2]*, Timothy Randolph*[2] *and Ali Shojaie*[3]
[1]Arizona State University
[2]Fred Hutch Cancer Research Center
[3]University of Washington
ywan1282@asu.edu

Diagnosis and treatment of human diseases require joint interpretation of molecular variations at multiple levels, often called multi-omics data. Recent advances in high-throughput sequencing technology have generated such multi-omics data from large cohorts of samples, providing unique opportunities to discover novel associations between biological levels and eventually build elaborate markers of disease. Several unsupervised joint-matrix-decomposition tools have been developed for extracting common and individual biological variations from multi-omics data, but few of them can examine how the extracted biological variations are associated with human diseases. This work addresses this limitation by introducing a novel high-dimensional regression framework to study the joint association between multi-omics data and a disease outcome. The proposed regression framework can be seamlessly used with a broad class of unsupervised joint-matrix-decomposition tools, and it allows both continuous and discrete outcomes. We demonstrate the effectiveness and efficiency of the proposed method using the multi-omics data collected in the "Carbohydrates and Related Biomarkers" study.

A Decomposition-based Canonical Correlation Analysis for High-Dimensional Datasets

♦*Hai Shu*[1]*, Xiao Wang*[2] *and Hongtu Zhu*[3]
[1]New York University
[2]Purdue University
[3]The University of North Carolina at Chapel Hill
hs120@nyu.edu

A typical approach to the joint analysis of two high-dimensional datasets is to decompose each data matrix into three parts: a low-rank common matrix that captures the shared information across datasets, a low-rank distinctive matrix that characterizes the individual information within a single dataset, and an additive noise matrix. Existing decomposition methods often focus on the orthogonality between the common and distinctive matrices, but inadequately consider the more necessary orthogonal relationship between the two distinctive matrices. The latter guarantees that no more shared information is extractable from the distinctive matrices. We propose decomposition-based canonical correlation analysis (D-CCA), a novel decomposition method that defines the common and distinctive matrices from the L2 space of random variables rather than the conventionally used Euclidean space, with a careful construction of the orthogonal relationship between distinctive matrices. D-CCA represents a natural generalization of the traditional canonical correlation analysis. The proposed estimators of common and distinctive matrices are shown to be consistent and have reasonably better performance than some state of-the-art methods in both simulated data and the real data analysis of breast cancer data obtained from The Cancer Genome Atlas.

Session 35: Modern streaming Data Analysis: process monitoring

Detection and identification

Olympia Hadjiliadis
Hunter College
olympia.hadjiliadis@gmail.com

We consider the problem of quickest detection of an abrupt change when there is uncertainty about the post-change distribution. In particular, we examine this problem in the continuous-time Wiener model where the drift of observations changes from zero to a random drift with a prescribed discrete distribution. We set up the problem as a stochastic optimization in which the objective is to minimize a measure of detection delay subject to a constraint on frequency of false alarms. We design a novel composite stopping rule and prove that it is asymptotically optimal of third order under a weighted Lorden criterion for detection delay. We also analyze the conditional identification error for the post-change drift asymptotically. Our composite rules are based on CUSUM stopping times, as well as their reaction periods, namely the times between the last reset of the CUSUM statistic process and the CUSUM alarm. The established results shed new light on the performance of CUSUM strategies under model uncertainty and offer new asymptotic optimality results in this framework.

Hot-spot detection and identification for Poisson data

Yujie Zhao, Xiaoming Huo and ♦*Yajun Mei*
Georgia Institute of Technology
ymei@isye.gatech.edu

In many bio-surveillance and healthcare applications, count data is very common to see, and they always come from different sources and are measured from many spatial locations repeatedly over time, say, daily/weekly/monthly. One of these count data is the number of patients get infected by some types of disease. In these applications, we are typically interested in detecting hot-spots in the infective rate, which are defined as some structured outliers that are sparse over the spatial domain. In this talk, we propose a method called "Poisson assisted Smooth Sparse Tensor Decomposition (PoSSTenD)", which can both detect when and localize where the hot-spots occur. The main idea of our proposed PoSSTenD method is articulated as follow. First, we represent the observed count data as a three-dimensional tensor including (i) a spatial dimension for location patterns, (ii) a category dimension for different types of data sources, (iii) a temporal domain for time patterns. Then we fit this tensor into a Poisson regression model, and then we further decompose the infective rate into two components: smooth global trend, local hot-spots. Next, we detect when the hot-spots occur by building a cumulative sum (CUSUM) control chart and localize where the hot-spots occur by the non-zero entries in the hot-spots estimations. The usefulness of our proposed methodology is validated through numerical simulation studies and a real-world dataset, which records the annual number of 10 different disease from 1993 to 2018 for 49 mainland states in the United States.

Adaptive Partially-Observed Sequential Change Point Detection for Covid-19 hotspots detection

Jiuyun Hu[1]*,* ♦*Hao Yan*[1]*, Yanjun Mei*[2] *and Sarah Holte*[3]
[1]Arizona State University
[2]Georgia Institution of Technology
[3]University of Washington
haoyan@asu.edu

Since the initial outbreak of the novel coronavirus in early January 2020, the COVID-19 pandemic has rapidly spread across the world. One question regarding this outbreak is to detect the outbreak in the future, then allocate proper resources and take measures to prevent the virus from further spread. The challenge in the real world is that the tests are limited and need to be distributed properly to different regions. The algorithm consists of three parts. Bayesian weighted

updated is used to get the posterior distribution of test statistics, Upper Confidence Bounds (UCB) is used to get the optimal distribution of tests for the next day, and CUSUM statistics is used to raise an alarm. The statistics is used for each region independently to make the decision of an alarm depending on whether it exceeds some threshold. The hotspots consist of all the regions where an alarm is raised. We conducted a simulation to get the threshold of the statistics to raise an alarm, and meanwhile compare the performance with a benchmark, even distribution across all regions. The threshold of the statistics of the algorithm and the benchmark is set to make the average run length in control around 430. The detection precision of the algorithm is similar to that of the benchmark, while the detection delay when out of control is smaller. The authors also observed the exploration and exploitation property of the algorithm. In the simulation work, we found that the distributions of different regions in the in control situation are similar, which is a good behavior of exploration. When out of control, the number of tests distributed in the out of control region dominated the other regions in most of the times, while the distributions of other regions are still similar. This shows a good behavior of both exploration and exploitation. In real case study, we analyzed the data in Washington State. The regions are 39 counties in WA. The data is daily county level positive COVID-19 cases in different counties. Yakima county was identified to be the first to raise an alarm. The hotspot regions consist of Adams county, Benton county, Douglas county, Franklin county Grant county and Yakima county.

Efficient Global Monitoring Statistics for High-Dimensional Data

Jun Li
University of California, Riverside
jun.li@ucr.edu

Global monitoring statistics play an important role in developing efficient monitoring schemes for high-dimensional data. A number of global monitoring statistics have been proposed in the literature. However, most of them only work for certain types of abnormal scenarios under specific model assumptions. How to develop global monitoring statistics that are powerful for any abnormal scenarios under flexible model assumptions is a long-standing problem in the statistical process monitoring field. To provide a potential solution to this problem, we propose a novel class of global monitoring statistics. Our proposed global monitoring statistics are easy to calculate and can work under flexible model assumptions since they can be built on any local monitoring statistic that is suitable for monitoring a single data stream. Our simulation studies show that the proposed global monitoring statistics perform well across a broad range of settings, and compare favorably with existing methods.

Session 36: Advances in Spatio-Temporal Statistics

Second-order semi-parametric inference for multivariate log Gaussian Cox processes

Kristian Hessellund[1], ♦*Ganggang Xu*[2], *Yongtao Guan*[2] *and Rasmus Waagepetersen*[1]
[1]Aalborg University
[2]University of Miami
gangxu@bus.miami.edu

This paper introduces a new approach to inferring the second-order properties of a multivariate log Gaussian Cox process (LGCP) with a complex intensity function. We assume a semi-parametric model for the multivariate intensity function containing an unspecified complex factor common to all types of points. Given this model, we exploit the availability of several types of points to construct a second-order conditional composite likelihood to infer the pair correlation and cross pair correlation functions of the LGCP. Crucially this likelihood does not depend on the unspecified part of the intensity function. We also introduce a cross-validation method for model selection and an algorithm for the regularized inference that can be used to obtain sparse models for cross-pair correlation functions. The methodology is applied to simulated data as well as data examples from microscopy and criminology. This shows how the new approach outperforms existing alternatives where the intensity functions are estimated non-parametrically.

Skew-Elliptical Cluster Processes

♦*Ngoc Anh Dao*[1] *and Marc Genton*[2]
[1]Texas A&M University
[2]King Abdullah University of Science and Technology
ngocanh@spruenker.de

The paper introduces skew-elliptical cluster processes. In contrast to the simple Gaussian isotropic structure of the distribution of the "children" events of a Thomas process, we propose an anisotropic structure by allowing the choice of a flexible covariance matrix and incorporating skewness or ellipticity parameters into the structure. Since the theoretical pair correlation functions of these processes are complex and analytically incomplete, and therefore the estimation of the parameters is computationally intensive, we propose reasonable approximations of the theoretical pair correlation functions of these cluster processes, which allow for a simpler parameter estimation. We present the estimation of their parameters using the minimum contrast method. For a data application, we use a fraction of the full redwood dataset. Our analysis shows that an elliptical cluster process can describe this point pattern better than a common Thomas process, since it is able to statistically model the non-circular shapes of the clusters in the data. The skew- elliptical cluster processes can be very meaningful for analyzing complex datasets in the field of spatial point processes since they provide more flexibility to detect interesting characteristics of the data.

Functional singular spectrum analysis

Hossein Haghbin[1], *S. Morteza Najibi*[2], *Rahim Mahmoudvand*[3], *Jordan Trinka*[4] *and* ♦*Mehdi Maadooliat*[5]
[1]Persian Gulf University
[2]Lund University
[3]Bu-Ali Sina University
[4]Pacific Northwest National Laboratory
[5]Marquette University
mehdi@mscs.mu.edu

This paper will develop a new extension of the singular spectrum analysis (SSA) called functional SSA to analyze functional time series. The new methodology is constructed by integrating ideas from functional data analysis and univariate SSA. Specifically, we introduce a trajectory operator in the functional world, which is equivalent to the trajectory matrix in the regular SSA. In the regular SSA, one needs to obtain the singular value decomposition (SVD) of the trajectory matrix to decompose a given time series. Since there is no procedure to extract the functional SVD (fSVD) of the trajectory operator, we introduce a computationally tractable algorithm to obtain the fSVD components. The effectiveness of the proposed approach is illustrated by an interesting example of remote sensing data. Also, we develop an efficient and user-friendly R package and a shiny web application to allow interactive exploration of the results.

Bayesian Co-kriging Model for Remote Sensing Measurements with Different Quality Flags: Uncertainty Quantification in NASA's AIRS Mission

Bledar Konomi
University of Cincinnati
KONOMIBR@ucmail.uc.edu

Motivated by different quality flag 3D measurements within a remote sensing instrument, this article explores a co-kriging model with separable structure Gaussian processes. Within this model, we develop an efficient Markov chain Monte Carlo (MCMC) algorithm that can be executed without storing or decomposing large matrices and direct simulation from conditional distributions. Moreover, we propose a computationally efficient recursively prediction procedure to make inference at a new location and atmospheric pressure. The methodological developments and statistical computing strategies that make this approach efficient are described in details. We apply the proposed method in the temperature measurements of a granule produced by the Atmospheric Infrared Sounder (AIRS) instrument, on board NASA's Aqua satellite, and compare its prediction performance and computational efficiency with other approaches.

Session 37: Advances in Spatial and Spatio-temporal Modeling and its Applications

A combined physical-statistical approach for estimating storm surge risk

♦*Whitney Huang*[1] *and Emily Tidwell*[2]
[1]Clemson University
[2]Dynetics
wkhuang@g.clemson.edu

Storm surge is an abnormal rise of seawater caused by a storm. According to the National Hurricane Center, storm surge is often the most damaging part of a hurricane. It poses the most severe threat to property and life in a coastal region. Thus, it is crucially important to assess the storm surge risk, typically summarized by r-year surge return level with return period r ranging from 10, 50, 100, or even much longer along a coastline. However, it is challenging to reliably estimate this quantity due to the limited storm surge observations in space and time. This talk presents an approach to integrate physical and statistical models to estimate extreme storm surge. Specifically, A physically-based hydrodynamics model is used to provide the needed interpolation in space and extrapolation in both time and atmospheric conditions. Statistical modeling is needed to 1) estimate the input distribution for running the computer model, 2) develop a statistical emulator in place of the computer simulator, and 3) estimate uncertainty due to input distribution, statistical emulator, missing/unresolved physics.

Scalable Forward Sampler Backward Smoother based on the Vecchia approximation

♦*Marcin Jurek*[1]*, Matthias Katzfuss*[2] *and Pulong Ma*[3]
[1]University of Texas at Austin
[2]Texas A&M University
[3]Duke University
marcin.jurek@austin.utexas.edu

We propose an approximation to the Forward Filter Backward (FFBS) sampler commonly used in Bayesian statistics when working with linear Gaussian state-space models. Traditional FFBS has proved immensely useful and fast but it requires inverting covariance matrices that have the size of the latent state vector. The computational burden associated with this operation effectively prohibits its applications in high-dimensional settings. In this paper we propose an approach based on the hierarchical Vecchia approximation of Gaussian processes, successfully used in spatial statistics. We extend the Vecchia approximation to variables without a spatial reference and use correlation distance instead. This allows us to apply the scalable version of the FFBS to any type of Gaussian data.

Conjugate spatio-temporal Bayesian multinomial Polya-gamma regression for the reconstruction of climate using pollen

John Tipton
1700 N Old Wire Rd
jrtipton@uark.edu

One of the most-widely available climate proxy data are tree pollen collected in sediments. Pollen grains in sediments are counted and the relative abundance of different tree species is a function of the underlying climate state. Thus, reconstructing patio-temporally correlated climate from pollen involves estimating a complex, non-linear relationship from multinomial data making traditional Markov Chain Monte Carlo methods difficult. In this work, I apply a Polya-gamma data augmentation scheme to enable conjugate parameter updates and reduce computational costs, allowing for Bayesian paleoclimate reconstructions from pollen to be performed at regional-to-continental scales.

Hierarchical Integrated Spatial Process Modeling of Monotone West Antarctic Snow Density Curves

♦*Philip White*[1]*, Durban Keeler*[2] *and Summer Rupper*[2]
[1]Brigham Young University
[2]University of Utah
pwhite@stat.byu.edu

Snow density estimates below the surface, used with airplane-acquired ice-penetrating radar measurements, give a site-specific history of snow water accumulation. Because it is infeasible to drill snow cores across all of Antarctica to measure snow density and because it is critical to understand how climatic changes are affecting the world's largest fresh water reservoir, we develop methods that enable snow density estimation with uncertainty in regions where snow cores have not been drilled. Snow density increases monotonically as a function of depth, except for possible micro-scale variability or measurement error, and it cannot exceed the density of ice. We present a novel class of integrated spatial process models that allow interpolation of monotone snow density curves. For computational feasibility, we construct the space-depth process through kernel convolutions of log-Gaussian spatial processes. We discuss model comparison, model fitting, and prediction. Using this model, we extend estimates of snow density beyond the depth of the original core and estimate snow density curves where snow cores have not been drilled. Along flight lines with ice-penetrating radar, we use interpolated snow density curves to estimate recent water accumulation and find predominantly decreasing water accumulation over recent decades.

Session 38: Recent Challenges and Advances for Complex Survival Data

Censored quantile regression with auxiliary information

Zhaozhi Fan
Memorial University
zhaozhi@mun.ca

Quantile regression models based on properly selected quantiles provide a global assessment of the covariate effects on the response.

In the analysis of survival data using quantile regression models, however, severe censoring could trigger problems such as the existence of an estimator of the regression coefficients for some extreme quantiles. There is a need of having samples with size as large as possible to have more event times included in the samples. In epidemiological studies, there is often times only a small portion of the whole study cohort that was accurately observed with some key exposures, the so-called validation sample, while for the rest of the cohort, only some inaccurate, auxiliary information was collected. In this talk we discuss a method that accommodates the non-validation sample into the modeling through a nonparametric smoothing technique and conduct quantile regression analysis based on the whole study cohort. The proposed estimator is consistent and has an asymptotic normal distribution. Simulation studies reveal that the proposed method is more efficient than the inference solely based on the validation sample. The method also provides us possibilities of looking into higher (lower) extreme quantiles of the failure distribution.

A comparative study of R packages for semiparametric shared frailty models

Tingxuan Wu[1]*, Longhai Li*[1] *and* ♦*Cindy Feng*[2]
[1]University of Saskatchewan
[2]cindy.feng@dal.ca
cindy.feng@dal.ca

Frailty models are often used to model the unobserved heterogeneity and clustered survival data. A shared frailty model is a random-effect model where the frailties are common or shared among individuals within groups. Different R packages are available for fitting shared frailty models, such as survival, frailtyEM, fraltypack, frailtysurv, and frailtyHL. However, little research has been conducted to compare the performance of various R packages for fitting shared frailty models, making it difficult for users to decide on an appropriate tool for analyzing clustered survival data. We aim to compare the performance of the R packages via a series of simulation studies. The bias and variance of the parameter estimates, rate of convergence, and computational time of the packages are compared. The advantages and limitations of the software are discussed in detail.

Accelerated Failure Time Models for Joint Analysis of Longitudinal and Time-to-Event Data

Shahedul Khan
University of Saskatchewan
khan@math.usask.ca

In joint modeling, the event time distribution depends on a longitudinally measured internal covariate. The proportional hazards (PH) family offers an attractive modeling paradigm for joint modeling. Although there are well known techniques to test the PH assumption for standard survival data analysis, checking this assumption for joint modeling has received less attention. An alternative framework involves considering an accelerated failure time (AFT) model, which is particularly useful when the PH assumption fails. Note that there are AFT models that can describe data with wide ranging characteristics but have received far less attention in survival data analysis. We develop methodology for joint modeling using the AFT family of distributions. Fitting a joint model is computationally and numerically much more demanding compared to standard survival data analysis. In particular, it requires to approximate multiple integrals that do not have analytic solutions except in very special cases. We propose computational algorithms for Bayesian inference, and develop a software package to fit these models. The proposed methodology is demonstrated using both simulated and real data.

Cox Proportional-Hazards Regression with Surrogates for Categorical Covariate and Left-Truncated Data

Hua Shen
University of Calgary
hua.shen@ucalgary.ca

Misclassification in categorical variables can often occur in medical research. Ignoring the misclassification issue in covariates when analyzing survival data often results in biased estimates of the regression coefficients and baseline hazard function. In the absence of a validation sample, latent class models can be adopted to address the issues. However, often the data is length biased. Restricting attention to the subjects that survived long enough to be in the sample result in biased estimates. We develop a likelihood-based approach in the framework of latent variables to address this challenge under the Cox proportional hazards model. Expectation-maximization algorithms are built up for both parametric and semiparametric model fitting. The performance of the proposed method is demonstrated in simulation studies and its application is illustrated on a breast cancer data.

Session 39: COVID-19 Modeling, Projection and Inference

Dynamic COVID risk assessment accounting for community virus exposure from a spatial-temporal transmission model

Yuanjia Wang
Columbia University
yw2016@cumc.columbia.edu

COVID-19 pandemic has caused unprecedented negative impacts on our society, including further exposing inequity and disparity in public health. To study the impact of socioeconomic factors on COVID transmission, we first propose a spatial-temporal model to examine the socioeconomic heterogeneity and spatial correlation of COVID-19 transmission at the community level. Second, to assess the individual risk of severe COVID-19 outcomes after a positive diagnosis, we propose a dynamic, varying-coefficient model that integrates individual-level risk factors from electronic health records (EHRs) with community-level risk factors. The underlying neighborhood prevalence of infections (both symptomatic and presymptomatic) predicted from the previous spatial-temporal model is included in the individual risk assessment so as to better capture the background risk of virus exposure for each individual. We design a weighting scheme to mitigate multiple selection biases inherited in EHRs of COVID patients. We analyze COVID transmission data in New York City (NYC, the epicenter of the first surge in the United States) and EHRs from NYC hospitals, where time-varying effects of community risk factors and significant interactions between individual- and community-level risk factors are detected. By examining the socioeconomic disparity of infection risks and interaction among the risk factors, our methods can assist public health decision-making and facilitate better clinical management of COVID patients.

Better Strategies for Containing COVID-19 Pandemic–A Study of 25 Countries via a vSIADR Model

Yumou Qiu
Iowa State University
yumouqiu@iastate.edu

We study epidemiological characteristics of 25 early COVID-19 outbreak countries, which emphasizes on the reproduction of infection and effects of government control measures. The study is based on a vSIADR model which allows asymptomatic and pre-diagnosis infections to reflect COVID-19 clinical realities, and a linear mixed-effect model to analyze the association between each country's control measures and the effective reproduction number. It finds significant effects of higher stringency measures in lowering the reproduction, and a significant shortening effect on the time to the epidemic turning point by applying stronger early countermeasures. Epidemic projections under scenarios of the counter measures (China and Korea, the US and the UK) show substantial reduction in the epidemic size and death by taking earlier and forceful actions. The governments' response before and after the start of the second wave epidemics were alarmingly weak, which made the average duration of the second wave more than doubled that of the first wave. We identify countries which urgently need to restore to at least the maximum stringency measures implemented so far in the pandemic in order to avoid even higher infection size and death.

Space-time Epidemic Models: The Complexity of Simplicity

Myungjin Kim[1]*, Zhiling Gu*[2]*, Shan Yu*[3]*,* ♦*Guannan Wang*[4] *and Li Wang*[5]
[1]Brown University
[2]Iowa State University
[3]University of Virginia
[4]College of William & Mary
[5]George Mason university
gwang01@wm.edu

Since the first case reported in December 2019, the outbreak of COVID-19 has expanded to touch nearly every corner of the world. To defend against COVID-19, it is essential to investigate how the SARS-CoV-2 virus spread, the significant factors that will affect the spread, as well as how they affect it and predict future development. When solving these problems, one crucial factor is the complexity of the model to describe the epidemic. Based on our study, it seems reasonable to choose models with different complexities during different pandemic periods. At the early stage of the pandemic, a simple model is preferred due to the sparsity of the cases. As the disease processes, a more complex model with a significant amount of flexibility can capture the heterogeneities and complexity of the underlying process. It is of great interest to balance the simplicity and flexibility of the models. In this work, we propose a flexible space-time epidemic model (STEM) to investigate the spatial-temporal pattern in the spread of COVID-19 at the area level. Within this modeling framework, we develop an automatic model identification method to adjust the complexity of the model by considering the significance and different types (i.e., varying or constant) of various factors. Moreover, we show that the estimators of constant coefficients and varying coefficient functions are consistent and asymptotically normal for constant coefficient estimators. The proposed method is evaluated by Monte Carlo simulation studies and applied to the COVID-19 analysis.

Analyzing COVID-19 Data: Some Issues and Challenges

Grace Yi
University of Western Ontario
gyi5@uwo.ca

The mystery of the coronavirus disease 2019 (COVID-19) and the lack of effective treatment for COVID-19 have presented a strikingly negative impact on public health. While research on COVID-19 has been ramping up rapidly, a very important yet somewhat overlooked challenge is on the quality and unique features of COVID-19 data. The manifestations of COVID-19 are not yet well understood. The swift spread of the virus is largely attributed to its stealthy transmissions in which infected patients may be asymptomatic or exhibit only flu-like symptoms in the early stage. Due to a good portion of asymptomatic infections, the confirmed cases are typically under-reported, error-contaminated, and involved with substantial noise. In this talk, I will discuss some issues related to faulty COVID-19 data and how they may challenge inferential procedures.

Session 40: Design and Analysis Experiments for Developing Mobile Health Devices

Building health application recommender system using partially penalized regression

♦*Eun Jeong Oh*[1]*, Min Qian*[2]*, Ken Cheung*[2] *and David Mohr*[3]
[1]University of Pennsylvania
[2]Columbia University
[3]Northwestern University
shanias@yonsei.ac.kr

Behavioral intervention technologies through mobile applications have a great potential to enhance patient care by improving ef¡U+FB01¿cacy. One approach to optimize the utility of mobile health applications is through an individualized health application recommender system. We formalize such a recommender system as a policy that maps individual subject information to a recommended app. We propose to estimate the optimal policy which maximizes the expected utility by partial regularization via orthogonality using the adaptive Lasso (PRO-aLasso). We also derive the convergence rate of the expected outcome of the estimated policy to that of the true optimal policy. The PRO-aLasso estimators are shown to enjoy the same oracle properties as the adaptive Lasso. Simulations studies show that the PRO-aLasso yields simple, more stable policies with better results as compared to the adaptive Lasso and other competing methods. The performance of our method is demonstrated through an illustrative example using IntelliCare mobile health applications.

A Zero-Inflated Mixture Model for Multivariate Data Clustering

Bin Cheng, ♦*Xiaoqi Lu and Ying-Kuen Cheung*
Columbia University
lx2170@caa.columbia.edu

Clustering is a common unsupervised learning method that helps to reveal hidden structures in data by grouping similar objects. However, such method has not been widely used in healthcare data whose distributions are usually zero-inflated. We proposed a parametric modeling approach with a two-layer hierarchical structure: the first layer models the zero-inflation pattern, while the second layer models the conditional distribution of the positive entries. Parameters are estimated by a regularized maximum likelihood estimation (MLE), using expectation-maximization (EM) algorithm.

Apply adaptive randomization to sequential multiple assignment randomized trial to develop a smartphone apps platform for depression management

Ying Kuen Cheung[1]*,* ♦*Xiaobo Zhong*[2] *and Bibhas Chakraborty*[3]
[1]Columbia University
[2]Bristol Myers Squibb
[3]National University of Singapore
tonizhong@gmail.com

Implementation study is an important tool for deploying state-of-the-art treatments from clinical efficacy studies into a treatment program, with dual goals of learning about the effectiveness of the treatment and improving the quality of care for patients enrolled in the program. Applying an adaptive randomization scheme in sequential multiple assignment randomized trial (SMART) can help incorporate historical data or opinions, include randomization for learning purposes, and improve patient care of the entire program, so as to effectively and efficiently select the best health care smartphone apps for individuals. We study the potential issues in applying adaptive randomization to SMART in an open-deployment setting to develop a smartphone platform for depression management apps recommendation, in which the candidates of apps introduced by the platform are a time-dependent variable. The Q-learning method was used for learning the outcome of apps selection. The results can benefit the healthcare apps developers and clinical trial investigators interested in developing smartphone apps for healthcare delivery.

Personalized Policy Learning using Longitudinal Mobile Health Data

♦*Xinyu Hu*[1]*, Min Qian*[2]*, Bin Cheng*[2] *and Ying Kuen Cheung*[2]
[1]Uber
[2]Columbia University
hxy719@gmail.com

We address the personalized policy learning problem using longitudinal mobile health application usage data. Personalized policy represents a paradigm shift from developing a single policy that may prescribe personalized decisions by tailoring. Specifically, we aim to develop the best policy, one per user, based on estimating random effects under generalized linear mixed model. With many random effects, we consider new estimation method and penalized objective to circumvent high-dimension integrals for marginal likelihood approximation. We establish consistency and optimality of our method with endogenous app usage. We apply our method to develop personalized push ("prompt") schedules in 294 app users, with a goal to maximize the prompt response rate given past app usage and other contextual factors. We found the best push schedule given the same covariates varied among the users, thus calling for personalized policies. Using the estimated personalized policies would have achieved a mean prompt response rate of 23% in these users at 16 weeks or later: this is a remarkable improvement on the observed rate (11%), while the literature suggests 3%-15% user engagement at 3 months after download. The proposed method compares favorably to existing estimation methods including using the R function "glmer" in a simulation study.

Session 41: Recent Advances in Statistical Inference for Discrete Structures

Global and Individualized Community Detection in Inhomogeneous Multilayer Networks

Shuxiao Chen[1]*, Sifan Liu*[2] *and* ♦*Zongming Ma*[1]
[1]University of Pennsylvania
[2]Stanford University
zongming@wharton.upenn.edu

In this talk, we discuss community detection in a stylized yet informative inhomogeneous multilayer network model. In our model, layers are generated by different stochastic block models, the community structures of which are (random) perturbations of a common global structure while the connecting probabilities in different layers are not related. Focusing on the symmetric two block case, we establish minimax rates for both global estimation of the common structure and individualized estimation of layer-wise community structures. Both minimax rates have sharp exponents. In addition, we provide an efficient algorithm that is simultaneously asymptotic minimax optimal for both estimation tasks under mild conditions. The optimal rates depend on the parity of the number of most informative layers, a phenomenon that is caused by inhomogeneity across layers.

Exact Clustering in Tensor Block Model: Statistical Optimality and Computational Limit

Anru Zhang
Duke University
anru.stat@gmail.com

High-order clustering aims to identify heterogeneous substructure in multiway dataset that arises commonly in neuroimaging, genomics, and social network studies. The non-convex and discontinuous nature of the problem poses significant challenges in both statistics and computation. In this talk, we propose a tensor block model and the computationally efficient methods, *high-order Lloyd algorithm* (HLloyd) and *high-order spectral clustering* (HSC), for high-order clustering in tensor block model. The convergence of the proposed procedure is established, and we show that our method achieves exact clustering under reasonable assumptions. We also give the complete characterization for the statistical-computational trade-off in high-order clustering based on three different signal-to-noise ratio regimes. Finally, we show the merits of the proposed procedures via extensive experiments on both synthetic and real datasets.

Hierarchical stochastic block model for community detection in multiplex networks

♦*Arash Amini*[1]*, Marina Paez*[2] *and Lizhen Lin*[3]
[1]UCLA
[2]Federal University of Rio de Janeiro
[3]University of Notre Dame
aaamini@ucla.edu

Multiplex networks have become increasingly more prevalent in many fields, and have emerged as a powerful tool for modeling the complexity of real networks. There is a critical need for developing inference models for multiplex networks that can take into account potential dependencies across different layers, particularly when the aim is community detection. We add to a limited literature by proposing a novel and efficient Bayesian model for community detection in multiplex networks. A key feature of our approach is the ability to model varying communities at different network layers. In contrast, many existing models assume the same communities for all layers. Moreover, our model automatically picks up the necessary number of communities at each layer (as validated by real data examples). This is appealing, since deciding the number of communities is a challenging aspect of community detection, and especially so in the multiplex setting, if one allows the communities to change across layers. Borrowing ideas from hierarchical Bayesian modeling, we use a hierarchical Dirichlet prior to model community labels across layers, allowing dependency in their structure. Given the community labels, a stochastic block model (SBM) is assumed for each layer. We develop an efficient slice sampler for sampling the posterior distribution of the community labels as well as the link probabilities between communities. In doing so, we address some unique challenges posed by coupling the complex likelihood of SBM with the hierarchical nature of the prior on the labels. An extensive empirical validation is performed on simulated

and real data, demonstrating the superior performance of the model over single-layer alternatives, as well as the ability to uncover interesting structures in real networks.

Maximum likelihood for high-noise group orbit estimation and single-particle cryo-EM

♦*Zhou Fan*[1]*, Roy R. Lederman*[1]*, Yi Sun*[2]*, Tianhao Wang*[1] *and Sheng Xu*[1]
[1]Yale University
[2]University of Chicago
zhou.fan@yale.edu

Motivated by applications to single-particle cryo-electron microscopy (cryo-EM), we study several problems of function estimation in a low SNR regime, where samples are observed under random rotations of the function domain. In a general framework of group orbit estimation with linear projection, we describe a stratification of the Fisher information eigenvalues according to a sequence of transcendence degrees in the invariant algebra, and relate critical points of the log-likelihood landscape to a sequence of method-of-moments optimization problems. This extends previous results for a discrete rotation group without projection. We then compute these transcendence degrees and the forms of these moment optimization problems for several examples of function estimation under SO(2) and SO(3) rotations, including a simplified model of cryo-EM as introduced by Bandeira, Blum-Smith, Kileel, Perry, Weed, and Wein. For several of these examples, we affirmatively resolve numerical conjectures that 3rd-order moments are sufficient to locally identify a generic signal up to its rotational orbit. For low-dimensional approximations of the electric potential maps of two small protein molecules, we empirically verify that the noise-scalings of the Fisher information eigenvalues conform with these theoretical predictions over a range of SNR, in a model of SO(3) rotations without projection.

Session 42: Advances in Detection of Change Points and Signals

Multiple Changepoint Detection for Time Series Data

Robert Lund
University of California, Santa Cruz
rolund@ucsc.edu

The problem of detecting multiple changepoints often arises with correlated data sequences (time series). This talk compares and contrasts some recent approaches to the problem, including penalized likelihood methods and techniques based on binary and wild binary segmentation. With penalized likelihoods, the best performing penalty is sought (AIC, BIC, MBIC, MDL, etc.); we also explore whether binary segmentation techniques can perform as well as penalized likelihood methods. In comparisons with multiple mean shifts, it is shown that binary segmentation methods are often easily fooled. Amongst penalized likelihood methods, we show that MBIC and MDL methods work well, that BIC methods work surprisingly well (the BIC penalty does not depend on where the changepoints lie), and that all procedures degrade as the autocorrelation in the series increases. A distance metric between two changepoint configurations with an arbitrary number of changepoints is developed to help aid in the comparisons.

Detection and estimation of local signals

♦*Xiao Fang*[1]*, Jian Li*[2] *and David Siegmund*[3]
[1]The Chinese University of Hong Kong
[2]Adobe
[3]Stanford University
xfang@sta.cuhk.edu.hk

To segment a sequence of independent random variables at an unknown number of change-points, we introduce new procedures that are based on thresholding the likelihood ratio statistic or the maximum score statistic. We derive analytic approximations for the probability of a false positive error when there are no change-points, and justifies some of the approximations rigorously. This is based on joint work with Jian Li and David Siegmund.

Minimax change point testing in the high-dimensional regression setting

Zifeng Zhao[1]*, Alessandro Rinaldo*[2] *and* ♦*Daren Wang*[1]
[1]Notre Dame
[2]CMU
dwang24@nd.edu

We study the problem of testing the presence of changes points in high-dimensional regression setting. We derive the null limiting distribution, alternative distribution and show that the power goes to 1. In addition, we show that our new approach achieve the optimal detection boundary even when p¿¿n.

Multiple Testing of Local Extrema for Detection of Change Points

♦*Dan Cheng*[1]*, Zhibing He*[1] *and Armin Schwartzman*[2]
[1]Arizona State University
[2]University of California San Diego
chengdan@asu.edu

A new approach to detect change points based on differential smoothing and multiple testing is presented forlong data sequences modeled as piecewise constant functions plus stationary ergodic Gaussian noise. As anapplication of the STEM algorithm for peak detection developed in Schwartzman et al. (2011) and Cheng and-Schwartzman (2017), the method detects change points as significant local maxima and minima after smoothingand differentiating the observed sequence. The algorithm, combined with the Benjamini-Hochberg procedurefor thresholding p-values, provides asymptotic strong control of the False Discovery Rate (FDR) and powerconsistency, as the length of the sequence and the size of the jumps get large. Simulations show that FDRlevels are maintained in non-asymptotic conditions and guide the choice of smoothing bandwidth.

Session 43: Recent development in high-dimensional statistics and machine learning

Statistical Inference Using Conformal Prediction

Jing Lei
Carnegie Mellon University
jinglei@andrew.cmu.edu

We consider the problem of testing the equality of the conditional distribution of a response variable given a set of covariates between two populations. Such a testing problem is related to transfer learning and causal inference. We develop a nonparametric procedure by combining recent advances in conformal prediction with some new ingredients such as a novel choice of conformity score and data-driven choices of weight and score functions. To our knowledge, this is the first successful attempt of using conformal prediction for testing statistical hypotheses beyond exchangeability. Our method is suitable for modern machine learning scenarios where the data has high dimensionality and large sample sizes, and can be effectively combined with existing classification algorithms to find

good weight and score functions. The performance of the proposed method is demonstrated in synthetic and real data examples.

Adaptive Estimation of Multivariate Regression with Hidden Variables

Yang Ning
Cornell
yn265@cornell.edu

This paper studies the estimation of the coefficient matrix θ in multivariate regression with hidden variables, $Y = (\theta)^T X + (B^*)^T Z + E$, where Y is a m-dimensional response vector, X is a p-dimensional vector of observable features, Z represents a K-dimensional vector of unobserved hidden variables, possibly correlated with X, and E is an independent error. The number of hidden variables K is unknown and both m and p are allowed but not required to grow with the sample size n. Since only Y and X are observable, we provide necessary conditions for the identifiability of θ. The same set of conditions are shown to be sufficient when the error E is homoscedastic. Our identifiability proof is constructive and leads to a novel and computationally efficient estimation algorithm, called HIVE. The first step of the algorithm is to estimate the best linear prediction of Y given X in which the unknown coefficient matrix exhibits an additive decomposition of θ and a dense matrix originated from the correlation between X and the hidden variable Z. Under the row sparsity assumption on θ, we propose to minimize a penalized least squares loss by regularizing θ via a group-lasso penalty and regularizing the dense matrix via a multivariate ridge penalty. Non-asymptotic deviation bounds of the in-sample prediction error are established. Our second step is to estimate the row space of B* by leveraging the covariance structure of the residual vector from the first step. In the last step, we remove the effect of hidden variable by projecting Y onto the complement of the estimated row space of B*. Non-asymptotic error bounds of our final estimator are established. The model identifiability, parameter estimation and statistical guarantees are further extended to the setting with heteroscedastic errors.

Augmented Direct Learning for Conditional Average Treatment Effect Estimation with Double Robustness

Haomiao Meng[1] *and* ♦*Xingye Qiao*[2]
[1]Amazon
[2]Binghamton University
qiao@math.binghamton.edu

Inferring the heterogeneous treatment effect is a fundamental problem in the sciences and commercial applications. In this paper, we focus on estimating Conditional Average Treatment Effect (CATE), that is, the difference in the conditional mean outcome between treatments given covariates. Traditionally, Q-Learning based approaches rely on the estimation of conditional mean outcome given treatment and covariates. However, they are subject to misspecification of the main effect model. Recently, simple and flexible one-step methods to directly learn (D-Learning) the CATE without model specifications have been proposed. However, these methods are not robust against misspecification of the propensity score model. We propose a new framework for CATE estimation, robust direct learning (RD-Learning), leading to doubly robust estimators of the treatment effect. The consistency for our CATE estimator is guaranteed if either the main effect model or the propensity score model is correctly specified. The framework can be used in both the binary and the multi-arm settings and is general enough to allow different function spaces and incorporate different generic learning algorithms. We conduct a thorough theoretical analysis of the prediction error of our CATE estimator using statistical learning theory under both linear and non-linear settings. The effectiveness of our proposed method is demonstrated by simulation studies and a real data example about an AIDS Clinical Trials study.

Efficient Learning of Optimal Individualized Treatment Rules

Weibin Mo and ♦*Yufeng Liu*
University of North Carolina at Chapel Hill
yfliu@email.unc.edu

Precision medicine is an emerging scientific topic for disease treatment and prevention. Given data with individual covariates, treatments and outcomes, researchers can search for the optimal individualized treatment rule (ITR). Existing methods typically require initial estimation of some nuisance models. The double robustness property that can protect from misspecification of either the treatment-free effect or the propensity score has been widely advocated. However, when model misspecification exists, a doubly robust estimate can be consistent but may suffer from downgraded efficiency. Furthermore, most existing methods do not account for the potential problem when the variance of outcome is heterogeneous. Such heteroscedasticity can greatly affect the estimation efficiency of the optimal ITR. In this talk, we will demonstrate that the consequences of misspecified treatment-free effect and heteroscedasticity can be unified as a covariate-treatment dependent variance of residuals. To improve efficiency of the estimated ITR, we propose an Efficient Learning (E-Learning) framework for finding an optimal ITR in the multi-armed treatment setting.

Session 44: Historical Controls for Medical Product Development

Historical Borrowing in Pediatric Clinical Trials: An Overview and Regulatory Perspective

James Travis
NA
james.travis@fda.hhs.gov

Pediatric drug development is typically more challenging than adult drug development. Some main challenges include relatively smaller patient populations, less willingness to expose children to clinical trials, particularly where placebos are used. Extrapolation of efficacy has been used in pediatrics1 in an all-or-nothing approach. That is, either no efficacy studies are needed if the disease and response to treatment are considered similar enough to adults, otherwise or if there is enough uncertainly then adequately powered clinical efficacy studies are needed. Even these extra options, pediatric drug development remains challenging. Statistical methods that utilize data from the previously conducted studies, such as Bayesian methods2 with informative priors or comparable frequentist methods have been proposed as options to reduce the burden of the pediatric studies on the pediatric patients. In this presentation we will give a brief overview of these methods used, discuss some recent case examples and discuss the regulatory perspective on their use. 1 Sun, H., Temeck, J. W., Chambers, W., Perkins, G., Bonnel, R., & Murphy, D. (2017). Extrapolation of Efficacy in Pediatric Drug Development and Evidence-based Medicine: Progress and Lessons Learned. Therapeutic innovation & regulatory science, 2017, 1-7. https://doi.org/10.1177/2168479017725558 2 Goodman, S. N., & Sladky, J. T. (2005). A Bayesian approach to randomized controlled trials in children utilizing information from adults: the case of Guillain-Barre. Clinical Trials, 2(4), 305-310. https://doi.org/10.1191/1740774505cn102oa

Data-Adaptive Weighting of Real-World and Randomized Controls Using Propensity Scores: Creating a Hybrid Control Arm

♦*Joanna Harton, Rebecca Hubbard and Nandita Mitra*
University of Pennsylvania
jograce@pennmedicine.upenn.edu

Clinical trials with a hybrid control arm, a control arm constructed from a combination of randomized patients and real-world data on patients receiving usual care in standard clinical practice, have the potential to decrease the cost of randomized trials. However, due to stringent trial inclusion criteria and differences in care quality between trials and community practice, randomized control patients will likely have superior outcomes compared to their real-world counterparts. We propose a new method for analyses of trials with a hybrid control arm that efficiently controls bias and type I error. Under our proposed approach, each real-world subject is weighted by a function of the propensity score reflecting their similarity to the randomized controls while randomized subjects receive full weight. This weighting allows for real-world patients that more closely resemble randomized controls to have a larger contribution to the likelihood while dissimilar subjects are discounted. Estimates of the treatment effect are obtained via Bayesian inference. We compare our approach to existing approaches via simulations and apply these methods to a study using EHR data.

NA

Anthony J. Hatswell
NA
anthony.j.hatswell@gmail.com

External control arms (also known as synthetic controls, historical controls, real world data, and by other names) are becoming an increasingly important source to inform healthcare decision making. These can provide a snapshot of current clinical practice and outcomes. Their use is particularly common (but not exclusively) in the context of uncontrolled / single arm studies, where all participants have the same treatment. In this context they allow the estimation of counterfactual outcomes i.e. what would happen if a patient didn't receive the new intervention. In disease areas where patients will experience many lines of treatment, synthetic control arms will most likely include data at different lines of therapy, which is unlikely to match the uncontrolled study. In this case there is little guidance in the literature about which line to choose as the 'index' point when creating an external control arm ('time zero', from which outcomes like survival are calculated). We present a simulation study where multiple methods (such as first line in, last line in, propensity scoring, and random line) are tested for bias and accuracy - with some failing on all counts.

Session 45: New machine learning methods using subgrouping

Dynamic tensor recommender systems

Yanqing Zhang[1], ♦*Xuan Bi*[2], *Niansheng Tang*[1] *and Annie Qu*[3]
[1]Yunnan University
[2]University of Minnesota
[3]University of California, Irvine
xbi@umn.edu

Recommender systems have been extensively used by the entertainment industry, business marketing and the biomedical industry. In addition to its capacity of providing preference-based recommendations as an unsupervised learning methodology, it has been also proven useful in sales forecasting, product introduction and other production related businesses. Since some consumers and companies need a recommendation or prediction for future budget, labor and supply chain coordination, dynamic recommender systems for precise forecasting have become extremely necessary. In this work, we propose a new recommendation method, namely the dynamic tensor recommender system, which aims particularly at forecasting future recommendation. The proposed method utilizes a tensor-valued function of time to integrate time and contextual information, and creates a time-varying coefficient model for temporal tensor factorization through a polynomial spline approximation. Major advantages of the proposed method include competitive future recommendation predictions and effective prediction interval estimations. In theory, we establish the convergence rate of the proposed tensor factorization and asymptotic normality of the spline coefficient estimator. The proposed method is applied to simulations, IRI marketing data and Last.fm data. Numerical studies demonstrate that the proposed method outperforms existing methods in terms of future time forecasting.

Crowdsourcing Utilizing Subgroup Structure of Latent Factor Modeling

♦*Qi Xu*[1], *Yubai Yuan*[1], *Junhui Wang*[2] *and Annie Qu*[1]
[1]UC Irvine
[2]City University of Hong Kong
qxu6@uci.edu

Crowdsourcing has emerged as an alternative solution for collecting large scale labels. However, the majority of recruited workers are not domain experts so their contributed labels could be noisy. In this paper, we propose a two-stage model to predict the true labels for binary and multicategory classification tasks in crowdsourcing. In the first stage, we fit the observed labels with a latent factor model and incorporate subgroup structures for both tasks and workers through a multi-centroid grouping penalty. A group-specific rotation is introduced to align workers with different task categories to solve the multicategory crowdsourcing tasks. In the second stage, we propose an angle-based approach to identify high-quality worker subgroups which are relied upon for assigning labels to tasks. As a theoretical contribution, we show the estimation consistency of latent factors and the prediction consistency of the proposed method. The simulation studies show that the proposed method outperforms the existing competitive methods assuming the subgroup structures within tasks and workers. We also demonstrate the application of the proposed method to real world problems and show its superiority.

A Tensor Factorization Recommender System with Dependency

♦*Jiuchen Zhang*[1], *Yubai Yuan*[2] *and Annie Qu*[3]
[1]PhD student
[2]PhD
[3]Chancellor's Professor
jiuchez@uci.edu

Dependency structure in recommender systems has been widely adopted in recent years to improve the prediction accuracy and potentially address the 'cold-start' problem. In this paper, we propose an innovative tensor-based recommender system, namely, the Tensor Factorization with Dependency (TFD). The proposed method utilizes shared factors to characterize the dependency between different modes, in addition to pairwise interaction tensor factorization to integrate information among multiple modes. One major advantage of the proposed method is to utilize the dependency structure among modes instead of across subjects dependency. Specifically, the proposed method provides flexibility for different dependency structure by incorporating different shared latent factor. In compu-

tation, we achieve scalable computation for handling scarse tensor with high missing rate. Our numerical studies demonstrate that the proposed method outperforms the existing methods, especially on prediction accuracy.

Session 46: Recent advances in the analysis of complex event time data

Semiparametric regression analysis of bivariate censored events in a family study of Alzheimer's disease

♦*Fei Gao*[1]*, Donglin Zeng*[2] *and Yuanjia Wang*[3]
[1]Fred Hutchinson Cancer Research Center
[2]University of North Carolina at Chapel Hill
[3]Columbia University
fgao@fredhutch.org

Assessing disease comorbidity patterns in families represents the first step in gene mapping for diseases and is central to the practice of precision medicine. One way to evaluate the relative contributions of genetic risk factor and environmental determinants of a complex trait (e.g., Alzheimer's disease [AD]) and its comorbidities (e.g., cardiovascular diseases [CVD]) is through familial studies, where an initial cohort of subjects are recruited, genotyped for specific loci, and interviewed to provide extensive disease history in family members. Because of the retrospective nature of obtaining disease phenotypes in family members, the exact time of disease onset may not be available such that current status data or interval-censored data are observed. All existing methods for analyzing these family study data assume single event subject to right-censoring so are not applicable. In this article, we propose a semiparametric regression model for the family history data that assumes a family-specific random effect and individual random effects to account for the dependence due to shared environmental exposures and unobserved genetic relatedness, respectively. To incorporate multiple events, we jointly model the onset of the primary disease of interest and a secondary disease outcome that is subject to interval-censoring. We propose nonparametric maximum likelihood estimation and develop a stable EM algorithm for computation. We establish the asymptotic properties of the resulting estimators and examine the performance of the proposed methods through simulation studies. Our application to a real world study reveals that the main contribution of comorbidity between AD and CVD is due to genetic factors instead of environmental factors.

Censored Linear Regression in the Presence or Absence of Auxiliary Survival Information

Ying Sheng
University of California at San Francisco
ying.sheng@ucsf.edu

There has been a rising interest in better exploiting auxiliary summary information from large databases in the analysis of smaller-scale studies which collect more comprehensive patient-level information. The purpose of this paper is twofold: firstly, we propose a novel approach to synthesize information from both the aggregate summary statistics and the individual-level data in censored linear regression. We show that the auxiliary information amounts to a system of non-smooth estimating equations and thus can be combined with the conventional weighted log-rank estimating equations by using the generalized method of moments (GMM) approach. The proposed methodology can be further extended to account for the potential inconsistency in information from different sources. Secondly, in the absence of auxiliary information, we propose to improve estimation efficiency by combining the overidentified weighted log-rank estimating equations with different weight functions via the GMM framework. To deal with the non-smooth GMM-type objective functions, we develop an asymptotics-guided algorithm for parameter and variance estimation. We establish the asymptotic normality of the proposed GMM-type estimators. Simulation studies show that the proposed estimators can yield substantial efficiency gain over the conventional weighted log-rank estimators. The proposed methods are applied to a pancreatic cancer study for illustration. This is joint work with Yifei Sun, Detian Deng, and Chiung-Yu Huang.

Recurrent event trees and ensembles

♦*Yifei Sun*[1]*, Sy Han (Steven) Chiou*[2] *and Chiung-Yu Huang*[3]
[1]Columbia University
[2]University of Texas at Dallas
[3]University of California San Francisco
ys3072@cumc.columbia.edu

Recurrent event data are commonly encountered in longitudinal studies. In this talk, we consider predicting the next occurrence of recurrent events using tree-based methods. We propose two types of approaches: event-specific prediction and global prediction for multiple or all events. Our methods allow the dependence structure among events to be completely unspecified and can easily incorporate previous event times as predictors. We also developed variable importance measures to gain an understanding of influential predictors. Extensive simulation studies and a data example are used to illustrate the proposed methods.

Statistical methods for semi-competing risks modeling with application to SEER-Medicare data

♦*Hong Zhu*[1]*, Yu Lan*[2]*, JIng Ning*[3] *and Yu Shen*[3]
[1]The University of Texas Southwestern Medical Center
[2]Southern Methodist University
[3]The University of Texas MD Anderson Cancer Center,
hong.zhu@utsouthwestern.edu

Semi-competing risks data often arise in medical studies where the terminal event (e.g., death) censors the non-terminal event (e.g., cancer recurrence), but the non-terminal event does not prevent the subsequent occurrence of the terminal event. In this talk, we consider regression modeling of semi-competing risks data to assess the covariate effects on the respective non-terminal and terminal event times. A copula-based framework for semi-competing risks regression with time-varying coefficients is proposed, where the dependence between the non-terminal and terminal event times is characterized by a copula and the time-varying covariate effects are imposed on two marginal regression models. A two-stage inferential procedure is developed for estimating the association parameter in the copula model and time-varying regression parameters. The finite sample performance of the proposed method is evaluated through simulation studies and the method is illustrated through an application to Surveillance, Epidemiology, and End Results-Medicare data for elderly women diagnosed with early-stage breast cancer and initially treated with breast-conserving surgery.

Session 47: Emerging Topics in Statistical Genetics and Genomics

Efficient SNP-based Heritability Estimation using Gaussian Predictive Process in Large-scale Cohort Studies

Souvik Seal[1]*, Abhirup Datta*[2] *and* ♦*Saonli Basu*[3]
[1]Colorado School of Public Health

[2]Johns Hopkins University
[3]University of Minnesota
saonli@umn.edu

For decades, linear mixed models (LMM) have been widely used to estimate heritability in twin and family studies. Recently, with the advent of high throughput genetic data, there have been attempts to estimate heritability from genome-wide SNP data on a cohort of distantly related individuals. However, fitting such an LMM in large-scale cohort studies is tremendously challenging due to high dimensional linear algebraic operations. In this paper, we simplify the LMM by unifying the concept of Genetic Coalescence and Gaussian Predictive Process, and thereby greatly alleviating the computational burden. Our proposed approach PredLMM has much better computational complexity than most of the existing packages and thus, provides an efficient alternative for estimating heritability in large-scale cohort studies. We illustrate our approach with extensive simulation studies and use it to estimate the heritability of multiple quantitative traits from the UK Biobank cohort.

Hurdle Poisson Model-based Clustering for Microbiome Data

Zhili Qiao and ♦*Peng Liu*
Iowa State University
PLIU@iastate.edu

High-throughput sequencing technologies have greatly facilitated microbiome studies. However, studying relationships among features is a challenging task given the sparsity and high-dimensionality of microbiome data. By grouping features with similar abundance profiles across treatments, cluster analysis provides insights into microbiome networks. In this presentation, we propose a model-based clustering algorithm based on Poisson hurdle models for sparse microbiome count data. We describe an expectation-maximization algorithm and a modified version using simulated annealing to conduct cluster analysis. We also provide methods for initialization and choosing the number of clusters. Simulation results demonstrate that our proposed methods provide better clustering results than alternative methods under a variety of settings. We also apply the proposed method to a sorghum rhizosphere microbiome dataset that results in interesting biological findings.

Sparsity in Single Cell Hi-C Data-Not All Zeros Are Created Equal

Shili Lin
Ohio State University
shili@stat.osu.edu

The prevalence of dropout events is a serious problem for single cell Hi-C data due to insufficient sequencing depth and data coverage, which brings difficulties in downstream studies such as clustering and structural analysis. Complicating things further is the fact that dropouts are confounded with structural zeros due to underlying biological mechanisms, leading to observed zeros being a mixture of both types of events. Although a great deal of progress has been made in imputing dropout events for single cell RNA-seq data, little has been done in identifying structural zeros and imputing dropouts for single cell Hi-C data. In this talk, I will first discuss the adaptation of several methods from the single cell RNA-seq literature for inference on observed zeros in single cell Hi-C data and the evaluation of their performance. I will then describe a Bayesian hierarchical model designed specifically for single cell Hi-C data to infer structural zeros, impute dropouts, and improve data quality in general. Simulation studies as well as real data applications will be described to illustrate the performance of the methods for imputing the zeros, and for clustering and structural analysis.

Model-based analysis of alternative polyadenylation using 3' end reads

♦*Wei Vivian Li*[1]*, Dinghai Zheng*[1]*, Ruijia Wang*[1] *and Bin Tian*[2]
[1]Rutgers, The State University of New Jersey
[2]Wistar Institute
wl522@sph.rutgers.edu

Most eukaryotic genes harbor multiple cleavage and polyadenylation sites (PASs), leading to expression of alternative polyadenylation (APA) isoforms. APA regulation has been implicated in a diverse array of physiological and pathological conditions. While RNA sequencing tools that generate reads containing the PAS, named onSite reads, have been instrumental in identifying PASs, they have not been widely used. By contrast, a growing number of methods generate reads that are close to the PAS, named nearSite reads, including the 3' end counting strategy commonly used in single cell analysis. How these nearSite reads can be used for APA analysis, however, is poorly studied. Here, we present a computational method, named model-based analysis of alternative polyadenylation using 3' end-linked reads (MAAPER), to examine APA using nearSite reads. MAAPER uses a probabilistic model to predict PASs for nearSite reads with high accuracy and sensitivity, and examines different types of APA events, including those in 3'UTRs and introns, with robust statistics. We show MAAPER's accuracy with data from both bulk and single cell RNA samples and its applicability in unpaired or paired experimental designs.

Session 48: Statistical considerations for master protocol in I-O and Cell Therapy

Statistical considerations for master protocol in I-O and Cell Therapy

Yuan Ji
The University of Chicago
yji@health.bsd.uchicago.edu

In modern I-O and Cell Therapy drug development, master protocol with adaptive designs including basket trial, umbrella trial or even platform study has been more and more utilized in the clinical development. It includes but not limited to dose determination on multiple indications, optimal combination or drug candidate selection, continuous monitoring for Go/No Go, subgroup enrichment with sample size re-estimation adaptive design, etc. This will often involve with historical or Bayesian borrowing, multiplicity adjustment, controlling and calibrating Statistical Errors and type I error control. In this session, we will have experts in this area to present their recent advanced work.

Statistical considerations for master protocol in I-O and Cell Therapy

Jianchang Lin[1]*, Yuan Ji*[2]*,* ♦*Peter Thall*[3] *and Shentu Yue*[4]
[1]Takeda Pharmaceuticals
[2]The University of Chicago
[3]The University of Texas MD Anderson Cancer Center
[4]Merck
rex@mdanderson.org

In modern I-O and Cell Therapy drug development, master protocol with adaptive designs including basket trial, umbrella trial or even platform study has been more and more utilized in the clinical development. It includes but not limited to dose determination on multiple indications, optimal combination or drug candidate selection, continuous monitoring for Go/No Go, subgroup enrichment with sample size re-estimation adaptive design, etc. This will often

involve with historical or Bayesian borrowing, multiplicity adjustment, controlling and calibrating Statistical Errors and type I error control. In this session, we will have experts in this area to present their recent advanced work.

Model-based inference with nonconcurrent control in Platform Trials

Anqi Pan[1], Nicole Li[2], Thomas Jemielita[2] and ♦Yue Shentu[2]
[1]School of Public Health, University of Georgia
[2]Merck & Co
yue_shentu@merck.com

Incorporation of nonconcurrent control may potential increase the efficiency of the platform trials. One concern in utilizing nonconcurrent control is the bias it could introduce due to time trend. Systematic difference in observed prognostic variables may explain some of the shift in outcome over time. We look at the covariate-adjusted analysis with or without explicit modeling of time trend.

Session 49: Recent Developments in Resampling Techniques

Asymptotics of cross-validation and the bootstrap.

Morgane Austern
Harvard
morgane.austern@gmail.com

Cross validation is a critical tool in evaluating the performance of machine learning and statistical models. However, despite its ubiquitous role, its theoretical properties are still not well understood. In this talk, we study the asymptotic properties of the cross-validated risk for a large class of models that satisfy stability conditions. We establish a central limit theorem and Berry-Esseen bounds, which enable us to compute asymptotically accurate confidence intervals and gain insight into the statistical speed-up of cross validation; some surprising behavior of cross-validation is discovered. When our stability conditions are not met, we observe a lack of universality in the limiting distribution of the cross-validated risk. Hence in general to be able to build consistent confidence intervals for the cross-validated risk another approach needs to be taken. In this talk we explore the bootstrap as such a method and establish general conditions under which it is asymptotically consistent for the cross validated risk, even when it fails to be asymptotically Gaussian.

Bootstrap-Assisted Inference for Generalized Grenander-type Estimators

♦*Matias Cattaneo[1], Michael Jansson[2] and Kenichi Nagasawa[3]*
[1]Princeton University
[2]UC-Berkeley
[3]University of Warwick
cattaneo@princeton.edu

Generalized Grenander-type estimators are an important class of monotone function estimators in statistics, econometrics, data science, and other areas. The generic limiting distribution of those estimators is characterized as the greatest convex minorant of a non-stationary Gaussian process, which depends on several nuisance functional parameters. Unfortunately, the standard non-parametric bootstrap is unable to consistently approximate the large sample distribution of the the generalized Grenander-type estimators, making statistical inference a challenging endeavor in applications. To solve this problem, we present a valid bootstrap-assisted inference procedure for the general class of generalized Grenander-type estimators, which relies on a carefully crafted and automatic transformation of the estimator. Our proposed method only requires the consistent estimation of a single scalar quantity, for which we propose an automatic procedure based on numerical derivative estimation. Under random sampling, our inference method restores validity of the exchangeable bootstrap, and therefore of the standard non-parametric bootstrap in particular. We illustrate our methods with several examples, and we also provide a small simulation study.

Dependence-Robust Inference Using Resampled Statistics

Michael Leung
UCSC
leungm@ucsc.edu

We develop inference procedures robust to general forms of weak dependence. The procedures utilize test statistics constructed by resampling in a manner that does not depend on the unknown correlation structure of the data. We prove that the statistics are asymptotically normal under the weak requirement that the target parameter can be consistently estimated at the parametric rate. This holds for regular estimators under many well-known forms of weak dependence and justifies the claim of dependence-robustness. We consider applications to settings with unknown or complicated forms of dependence, with various forms of network dependence as leading examples. We develop tests for both moment equalities and inequalities.

On Gaussian Approximation for M-Estimator

♦*Masaaki Imaizumi[1] and Taisuke Otsu[2]*
[1]The University of Tokyo
[2]London School of Economics
imaizumi@g.ecc.u-tokyo.ac.jp

This study develops a non-asymptotic Gaussian approximation theory for distributions of M-estimators, which are defined as maximizers of empirical criterion functions. In existing mathematical statistics literature, numerous studies have focused on approximating the distributions of the M-estimators for statistical inference. In contrast to the existing approaches, which mainly focus on limiting behaviors, this study employs a non-asymptotic approach, establishes abstract Gaussian approximation results for maximizers of empirical criteria, and proposes a Gaussian multiplier bootstrap approximation method. Our developments can be considered as extensions of the seminal works (Chernozhukov, Chetverikov and Kato (2013, 2014, 2015)) on the approximation theory for distributions of suprema of empirical processes toward their maximizers. Through this work, we shed new lights on the statistical theory of M-estimators. Our theory covers not only regular estimators, such as the least absolute deviations, but also some non-regular cases where it is difficult to derive or to approximate numerically the limiting distributions such as non-Donsker classes and cube root estimators.

Session 50: Modern techniques in statistical learning

Deep Neural Network with a Smooth Monotonic Output Layer for Dynamic Risk Prediction

♦*Zhiyang Zhou[1], Yu Deng[1], Lei Liu[2], Hongmei jiang[3], Yifan Peng[4], Xiaoyun Yang[1], Yun Zhao[5], Hongyan Ning[1], Norrina Allen[1], John Wilkins[1], Kiang Liu[1], Donald Lloyd-Jones[1] and Lihui Zhao[1]*
[1]Northwestern University Feinberg School of Medicine
[2]Washington University School of Medicine in St. Louis
[3]Northwestern University
[4]Weill Cornell Medicine
[5]University of California, Santa Barbara
zhiyang.zhou@northwestern.edu

Risk prediction is a key component of survival analysis in medicine, public health, economics, engineering, and many other areas. The fundamental concern of risk prediction is the relationship between predictors and the distribution of time to event (i.e., the survival function). The recent success of survival analysis has already been extended to dynamic risk prediction, where the model considers repeated measurements of time-varying predictors. However, existing approaches usually involve strong model assumptions (e.g., additive effects and/or proportional hazard) or discretize the time domain and approximate the survival function by a step function, which may lead to biased prediction. To tackle these issues, we present a deep neural network with a novel output layer termed the Smooth Monotonic Output Layer (SMOL). The resulting network involves no discretization and specifies no parametric structure for the underlying relationship between predictors and the time to event. At the core, SMOL takes a general vector as the input and constructs a monotonic and differentiable function via B-spline approximation. Attaching SMOL to a neural network, one may infer/learn the cumulative distribution function for a continuous random variable, directly and nonparametrically. We conduct experiments on datasets from the Lifetime Risk Pooling Project (LRPP). LRPP pools together individual data from twenty community-based studies on cardiovascular disease — the leading cause of death in the world — and involves around three hundred thousand participants with long-term follow-ups of longitudinal risk factors (e.g., blood pressure and cholesterol). Extensive results show that our proposal achieves state-of-the-art accuracy in predicting the individual-level risk of atherosclerotic cardiovascular disease.

Statistical Disaggregation — a Monte Carlo approach under Constraints

Shenggang Hu, ♦*Hongsheng Dai and Fanlin Meng*
University of Essex
hdaia@essex.ac.uk

Statistical disaggregation has become more and more important for smart energy systems. A typical example for such disaggregation problems is to learn energy consumption for a higher resolution level (data recorded at higher frequency) based on data at a lower resolution (data recorded at lower frequency). Constrained models are often used in such problems and they are often very useful comparing to their unconstrained counterparts in terms of reducing uncertainty and lead to an improvement of the overall performance. However, these constrained models usually are not expressible as ordinary distributions due to their intractable density functions which makes it hard to conduct further analysis. This talk will present a novel constrained Monte Carlo sampling algorithm based on Langevin diffusions and rejection sampling to solve the problem of sampling from constrained models. This new method is then applied to a statistical disaggregation problem for an electricity consumption dataset. Our approach provides excellent accuracy of data imputation, based on our simulation studies and data analysis.

Multimodal Neuroimaging Data Integration and Pathway Analysis

♦*Yi Zhao*[1]*, Lexin Li*[2] *and Brian Caffo*[3]
[1]Indiana University School of Medicine
[2]University of California Berkeley
[3]Johns Hopkins Bloomberg School of Public Health
yz125@iu.edu

With fast advancements in technologies, the collection of multiple types of measurements on a common setof subjects is becoming routine in science. Some notable examples include multimodal neuroimagingstudies for the simultaneous investigation of brain structure and function and multi-omics studies forcombining genetic and genomic information. Integrative analysis of multimodal data allows scientists tointerrogate new mechanistic questions. However, the data collection and generation of integrativehypotheses is outpacing available methodology for joint analysis of multimodal measurements. In thisarticle, we study high-dimensional multimodal data integration in the context of mediation analysis. We aimto understand the roles different data modalities play as possible mediators in the pathway between anexposure variable and an outcome. We propose a mediation model framework with two data types servingas separate sets of mediators and develop a penalized optimization approach for parameter estimation. Westudy both the theoretical properties of the estimator through an asymptotic analysis and its fi nite-sampleperformance through simulations. We illustrate our method with a multimodal brain pathway analysishaving both structural and functional connectivities as mediators in the association between sex andlanguage processing.

Disclosure control for microdata: a mixture modeling approach

♦*Bei Jiang*[1]*, Adrian Raftery*[2]*, Russel Steele*[3] *and Naisyin Wang*[4]
[1]University of Alberta
[2]University of Washington
[3]McGill University
[4]University of Michigan
bei1@ualberta.ca

The statistical disclosure control (SDC) methods is a class of privacy and utility preserving techniques that deliberately perturb the original data before public release. The goal of SDC methods is to reduce the disclosure risks to an acceptable level, while releasing public-use data sets (known as synthetic data sets) that still perfectly preserve the information from the original data set. In this work, we investigate a mixture-based multiple imputation synthetic method that provides different degrees of perturbation to records/individuals of different levels of disclosure risk. The first step of the method utilizes the concept of k-Anonymity proposed by Sweeney (2002) to divide individuals into subgroups of different disclosure risk levels, using the given risk thresholds. Then, through a data augmentation step, we introduce a tuning mechanism when building imputation models, to further control information loss and hence provide different levels of protection to individuals in different risk subgroups. We illustrate the proposed method using a real data application.

Session 51: Statistical Methods for Single-Cell Data

SnapHiC: a computational pipeline to identify chromatin loops from single cell Hi-C data

Ming Hu
Cleveland Clinic
hum@ccf.org

Single cell Hi-C (scHi-C) analysis has been increasingly used to map chromatin architecture in diverse tissue contexts, but computational tools to define chromatin loops at high resolution from scHi-C data are still lacking. Here, we describe single nucleus analysis pipeline for Hi-C (SnapHiC), a method that can identify chromatin loops at high resolution and accuracy from scHi-C data. Using scHi-C data from 742 mouse embryonic stem cells, we benchmark SnapHiC against a number of computational tools developed for mapping chromatin loops and interactions from bulk Hi-C. We further demonstrate its utility by analyzing single-nucleus methyl-3C-seq data from 2,869 human prefrontal cortical cells, which uncovers cell-type-specific chromatin loops and predicts putative target

genes for non-coding sequence variants associated with neuropsychiatric disorders. Our results suggest that SnapHiC could facilitate the analysis of cell-type-specific chromatin architecture and gene regulatory programs in complex tissues.

Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer

Chi-Yun Wu and ♦*Nancy Zhang*
University of Pennsylvania
nzh@wharton.upenn.edu

Cancer progression is driven by both somatic copy number aberrations (CNAs) and chromatin remodeling, yet little is known about the interplay between these two classes of events in shaping the clonal diversity of cancers. In this talk I will discuss the problem of allele-specific copy number estimation in single-cell DNA and/or ATAC sequencing data, and present our recently published method Alleloscope. Such analysis in scATAC-seq data enables the combined analysis of allele-specific copy number and chromatin accessibility. First, on scDNA-seq data from gastric, colorectal, and breast cancer samples, with validation using matched linked-read sequencing, Alleloscope finds pervasive occurrence of highly complex, multi-allelic copy number aberrations, where cells that carry varying allelic configurations adding to the same total copy number co-evolve within a tumor. On scATAC-seq from two basal cell carcinoma samples and a gastric cancer cell line, Alleloscope detects multi-allelic copy number events and copy neutral loss-of-heterozygosity, enabling dissection of the contributions of chromosomal instability and chromatin remodeling to tumor evolution.

Jointly Defining Cell States from Single-Cell Multi-Omic and Spatial Transcriptomic Datasets

Joshua Welch
University of Michigan
welchjd@umich.edu

Single-cell omic technologies provide an unprecedented opportunity to define molecular cell states in a data-driven fashion, but present unique data integration challenges. We developed an online learning algorithm for single-cell multi-omic integration, allowing highly scalable integration and the ability to incorporate new datasets without recalculating results from scratch. Furthermore, integration analyses often involve datasets with partially overlapping features, including both shared features that occur in all datasets and features exclusive to a single experiment. Previous computational integration approaches require that the input matrices share the same number of either genes or cells, and thus can use only shared features. To address this limitation, we developed a novel algorithm for "mosaic integration" of single-cell datasets containing both shared and unshared features. Additionally, new technologies for simultaneously profiling gene expression and epigenomic state in the same cells enable investigation of the correspondence between cell states inferred from different molecular layers. To realize this potential, we developed an approach for modeling epigenetic regulation of gene expression from single-cell multi-omic data, allowing us to quantify the degree of concordance or decoupling between transcriptomic and epigenomic states.

TWO-SIGMA-G: A New Competitive Gene Set Testing Framework for scRNA-seq Data

♦*Eric Van Buren*[1]*, Ming Hu*[2]*, Liang Cheng*[3]*, John Wrobel*[3]*, Kirk Wilhelmsen*[3]*, Lishan Su*[3]*, Yun Li*[3] *and Di Wu*[3]
[1]Harvard University
[2]Cleveland Clinic
[3]University of North Carolina at Chapel Hill
evb@hsph.harvard.edu

We propose TWO-SIGMA-G, a competitive gene set test designed for scRNA-seq data. TWO-SIGMA-G uses the mixed-effects regression modelling approach of our previously published TWO-SIGMA to test for differential expression at the gene-level. This regression-based approach allows TWO-SIGMA-G to accommodate zero-inflated and overdispersed counts, within-sample cell-cell correlation, and to fully analyze complex modern experimental designs, which often necessitate controlling for multiple confounding covariates in large samples of cells. TWO-SIGMA-G uses a novel approach to adjust for inter-gene-correlation (IGC) at the set-level, which can inflate type-I error when ignored. Simulations demonstrate that TWO-SIGMA-G preserves type-I error and increases power in the presence of IGC compared to other methods designed for bulk and single-cell RNA-seq data. Application to two real datasets of HIV infection in mice and Alzheimer's disease progression in humans reveal biologically meaningful results.

Session 52: New Challenges in Functional Data Analysis

A reproducing kernel Hilbert space framework for functional classification

♦*Peijun Sang*[1]*, Adam Kashlak*[2] *and Linglong Kong*[2]
[1]University of Waterloo
[2]University of Alberta
peijun.sang@uwaterloo.ca

The intrinsic infinite-dimensional nature of functional data creates a bottleneck in the application of traditional classifiers to functional settings. These classifiers are generally either unable to generalize to infinite dimensions or have poor performance due to the curse of dimensionality. To address this concern, we propose building a distance-weighted discrimination (DWD) classifier on scores obtained by projecting data onto one specific direction. We choose this direction by minimizing, over a reproducing kernel Hilbert space, an empirical risk function containing the DWD classifier loss function. Our proposed classifier avoids overfitting and enjoys the appealing properties of DWD classifiers. We further extend this framework to accommodate functional data classification problems where scalar covariates are involved. In contrast to previous work, we establish a non-asymptotic estimation error bound on the relative misclassification rate. Through simulation studies and a real-world application, we demonstrate that the proposed classifier performs favorably relative to other commonly used functional classifiers in terms of prediction accuracy in finite-sample settings.

Functional L-Optimality Subsampling for Massive Data

Hua Liu[1]*, Jinhong You*[1] *and* ♦*Jiguo Cao*[2]
[1]Shanghai University of Finance and Economics
[2]Simon Fraser University
jiguo_cao@sfu.ca

Massive data bring the big challenges of memory and computation for analysis. These challenges can be tackled by taking subsamples from the full data as a surrogate. For functional data, it is common to collect multiple measurements over their domains, which require even more memory and computation time when the sample size is large. The computation would be much more intensive when statistical inference is required through bootstrap samples. To the best of our knowledge, this article is the first attempt to study the subsampling method for the functional linear model. We propose an optimal subsampling method based on the functional L-optimality criterion. When the response is a discrete or categorical variable, we further extend our proposed functional L-optimality

subsampling (FLoS) method to the functional generalized linear model. We establish the asymptotic properties of the estimators by the FLoS method. The finite sample performance of our proposed FLoS method is investigated by extensive simulation studies. The FLoS method is further demonstrated by analyzing two large-scale datasets: the global climate data and the kidney transplant data. The analysis results on these data show that the FLoS method is much better than the uniform subsampling approach and can well approximate the results based on the full data while dramatically reducing the computation time and memory.

Nonparametric Estimation of Repeated Densities with Heterogeneous Sample Sizes

Jiaming Qiu, ♦*Xiongtao Dai and Zhengyuan Zhu*

Iowa State University

xdai@iastate.edu

We consider the estimation of densities in multiple subpopulations, where the available sample size in each subpopulation varies greatly. For example, in the context of epidemiology, the age distributions of patients with different conditions is of central interest, where a disease defines a subpopulation. A key challenge in estimating the age distributions comes from the highly variable sample sizes in the subpopulations, making the estimation especially difficult for rare conditions. We propose a fully data-driven approach to estimate the densities without specifying a parametric form of the density families. The idea is to map the density functions to a Hilbert space and then apply functional data analytic methods to derive low-dimensional approximates. Subpopulation densities are then fitted within the low-dimensional families using likelihood-based methods, where information borrowing is enforced through shrinkage. Application to electronic health record data will be showcased.

Optimal Imperfect Classification for Functional Data

♦*Shuoyang Wang*[1]*, Zuofeng Shang*[2]*, Guanqun Cao*[1] *and Jun Liu*[3]

[1]Auburn University

[2]New jersey Institute of Technology

[3]Harvard University

szw0100@auburn.edu

A central topic in functional data analysis is how to design an optimaldecision rule, based on training samples, to classify a data function. Thisproblem has been studied in the context of perfect classification in the sensethat Bayes risk asymptotically vanishes. In practical applications, Bayes riskis generally nonvanishing, called as imperfect classification, often resultingfrom the equivalent probability measures of the populations. Constructionof optimal classifiers in the latter setting becomes more challenging due tothe "closeness" of the populations. In this paper, we exploit the optimal im-perfect classification problem when data functions are Gaussian processes.Sharp nonasymptotic convergence rates for minimax excess misclassificationrisk are derived in both settings that data functions are fully observed anddiscretely observed. We propose two novel and easily implementable clas-sifiers based on discriminant analysis and deep neural network which areboth proven to achieve optimality. In discretely observed case, we discovera critical sampling frequency that governs the sharp convergence rates. Theproposed classifiers perform favorably in finite-sample applications, as wedemonstrate through comparisons with other functional classifiers in simula-tions and one real data application.

Session 53: Master Protocols - Theories, Applications, and When to Use It

FDA Review Perspectives on Master Protocols in Oncology

Erik Bloomquist

FDA

erik.bloomquist@fda.hhs.gov

Master protocols have become an attractive option in drug development, however there are many regulatory issues to consider when putting these designs into practice. This talk will focus on several important regulatory aspects on master protocol designs in oncology and highlight how master protocol designs still are not cookbook recipes. There remains the need to tailor master protocols to the specific setting under considerations. Topics from this talk will include single arm master protocols for drug discovery, false-positive error control, concurrent vs. non-control control arms, and adaptive features such as randomization.

Methodological Challenges in Collaborative Platform Trials

♦*Martin Posch, Marta Bofill Roig and Franz Konig*

Medical University of Vienna

martin.posch@meduniwien.ac.at

Platform trials are multi-armed trials where novel interventions can enter the platform over time if new treatments become available. On the other hand, treatment arms can leave the platform if a treatment arm completed recruitment or is stopped for futility or efficacy in an interim analysis. Platform trials offer the possibility to share control groups and thereby can improve the efficiency of drug development. Due to their adaptive nature, they provide sufficient flexibility to tailor important trial design aspects (as, e.g., sample sizes, specific inclusion, exclusion criteria) to the requirements of new compounds entering the platform [2]. The flexibility of platform trials, however, comes with challenges for statistical inference [1,3,4]. Issues to be addressed are the impact of adaptations (as the addition or dropping of arms has an impact on the statistical operating characteristics), multiplicity and the use of shared controls. Due to the staggered addition of treatments into the platform, sharing controls in platform trials comes with the additional complexity that some of the control subjects may have been randomised before a specific treatment arm entered the platform. Inclusion of such non-concurrent controls in statistical comparisons may introduce bias and several proposals for appropriate adjustments have been made. We will discuss the advantages and limitations of the different approaches with respect to power and type 1 error rate control. Furthermore, we give an overview on the current debate on multiplicity adjustment in platform trials, and discuss different approaches to account for the comparison of multiple treatment arms, multiple subgroups and interim analyses. [1] Collignon O., Gartner C., Haidich A.-B., Hemmings R.J., Hofner B., Pétavy F., Posch M., Rantell K., Roes K., Schiel A. Current Statistical Considerations and Regulatory Perspectives on the Planning of Confirmatory Basket, Umbrella, and Platform Trials. Clinical Pharmacology & Therapeutics 107(5), 1059-1067, (2020) [2] Collignon, O., Burman, C. F., Posch, M., & Schiel, A. (2021). Collaborative platform trials to fight COVID?19: methodological and regulatory considerations for a better societal outcome. Clinical Pharmacology & Therapeutics. [3] Meyer E.L., Mesenbrink P., Dunger-Baldauf C., Fülle H.-J., Glimm E., Li Y., Posch M., König F. The Evolution of Master Protocol Clinical Trial Designs: A Systematic Literature Review. Clinical Therapeutics 42(7), 1330-1360, (2020) [4] Posch, M., & König, F. (2020). Are p-values Useful to Judge the Evidence Against the Null Hypotheses in Complex Clinical Trials? A Comment on "The Role of p-values in Judging

the Strength of Evidence and Realistic Replication Expectations". Statistics in Biopharmaceutical Research, 1-3, (2002)

Simultaneous False-Regulatory Error Rates in Master protocols with Shared Control: False Discovery Rate Perspective

Jingjing Ye
BeiGene
jingjing.ye@beigene.com

Platform trial is a type of master protocol where multiple therapies can be investigated in the same trial. A shared control can be used for multiple therapies to gain operational efficiency and gain attraction to patients. To balance between controlling for false positive and having adequate power for detecting true signals, the impact of False Discovery Rate (FDR) is evaluated when multiple investigational drugs are studied in the master protocol. With the shared control group, the "random high" or "random low" in the control group can potentially impact all hypotheses testing that compare each of the test regimens and the control group in terms of probability of having at least one positive hypothesis outcome, or multiple positive hypotheses outcome. When regulatory agencies make the decision of approving or declining an application based on the master protocol design, this introduces a different type of error: simultaneous false-regulatory error. In this manuscript, we examine in details the derivations and properties of the simultaneous false-regulatory error in the master protocol with shared control under the framework of false discovery rate. The simultaneous false-regulatory error consists two parts: simultaneous false-discovery rate (SFDR) and simultaneous false non-discovery rate (SFNR). Based on our derivations and numerical analyses, strict error control to a pre-specified level on SFDR or SFNR, or reduce the type I error allocated to each individual treatment comparison to the shared control is deemed unnecessary.

Practical Considerations of Implementing Master Protocols for non-oncology studies in Industry

♦*Ling Wang, Pranab Ghosh and Satrajit Roychoudhury*
Pfizer
ling.wang2@pfizer.com

Master protocols, including platform trials, have been used in many settings, especially in trials sponsored by academic institutions in oncology. Less has been shared on implementing this approach in non-oncology setting done by industry sponsors. In this talk I will discuss practical considerations of running these studies in this setting, including statistical design, shared control arm, randomization and study operations. I will also share some simulation work on master protocols for non-oncology combination studies to improve efficiency compared to many separate protocols.

Virtual Poster & Mixer

Ambiguity Resolution to Enhance BERT Question-Answering

♦*Muzhe Guo*[1]*, Muhao Guo*[2]*, Edward Dougherty*[3] *and Fang Jin*[1]
[1]the George Washington University
[2]Arizona State University
[3]Roger Williams University
muzheguo@umich.edu

Bidirectional Encoder Representations from Transformers(BERT) models for question answering(QA) achieve impressive results on the SQuAD dataset, and so they are widely used for a variety of passage-based QA tasks. In some fields, especially the medical field, the content of paragraphs is complex and ambiguous, but higher accuracy is required. In such a situation, the BERT model for question answering performs not well. Therefore, strategies and techniques to enhance BERT QA models in the presence of complexity and ambiguity are highly-valuable. To address this issue, we introduce a novel approach called the Multiple Synonymous Questions BERT (MSQ-BERT) model for QA, which integrates question augmentation, rather than the typical single question used by the traditional BERT QA model, to elevate performance. We use Word Frequency Scores to highlight the score for each synonymous question and use Singular Value Decomposition(SVD) to reduce the rank of the score matrix to finally determine the answers to questions. Experiments with an ambiguous and complex dataset demonstrate a significant performance improvement of the MSQ-BERT method, demonstrating a new approach to addressing ambiguity in question-answering tasks.

An Interactive Knowledge Graph Based Platform for COVID-19Clinical Research

♦*Juntao Su*[1]*, Edward Dougherty*[2]*, Shuang Jiang*[1] *and Fang Jin*[1]
[1]George Washington University
[2]Roger Williams University
sujuntao@gwu.edu

Since the first identified case of COVID-19 in December 2019, a plethora of pharmaceuticals and therapeutics have been tested for COVID-19 treatment. While medical advancements and breakthroughs are well underway, the sheer number of studies, treatments, and associated reports makes it extremely challenging to keep track of the rapidly growing COVID-19 research landscape. While existing scientific literature search systems provide basic document retrieval, they fundamentally lack the ability to explore data, and in addition, do not help develop a deeper understanding of COVID-19 related clinical experiments and findings. As research expands, results do so as well, resulting in a position that is complicated and overwhelming. To address this issue, we present a named entity recognition based framework that accurately extracts COVID-19 related information from clinical test results articles, and generates an efficient and interactive visual knowledge graph. This knowledge graph platform is user friendly, and provides intuitive and convenient tools to explore and analyze COVID-19 research data and results including medicinal performances, side effects and target populations.

Machine-Understandable Saliency Estimation In Natural Language Processing

♦*Zhou Yang and Shunyan Luo*
George Washington University
zhou_yang@gwu.edu

Deep neural networks have achieved tremendous success and even surpassed human performance in various areas. However, the usage of deep learning models is largely limited in some tasks like automating medical diagnosis, resume screening, etc., where interpretability matters significantly. In the neural language processing domain, attention mechanism is treated as inherent explanation on deep NLP models. However, there is lots of buzz on its fidelity. As a result, we adopted linearly gradient gradient(LEG), a saliency estimation framework origin in computer vision and proposed a model-understandable saliency estimation(MUSE) by making normalized linear Gaussian perturbations on the encoding layer to interpret NLP models. A variant of MUSE with regularization is proposed to enhance interpretability to humans. The interpretability is verified through a crowd-sourced human evaluation study while sensitivity analysis is conducted to validate fidelity by interpreting LSTM and BERT model on three dataset among state-of-art interpretation

methods. MUSE achieves excellence performance on both studies. We also conduct sanity check to measure the sensitivity of MUSE.

Optimal Treatment Decision Rules in Precision Medicine based on Outcome Trajectories and Biosignatures

♦*Lanqiu Yao and Thaddeus Tarpey*
NYU School of Medicine
ly1192@nyu.edu

A pressing challenge in medical research is to identify optimal treatments for individual patients. This is particularly challenging in mental health settings where mean responses are often similar across multiple treatments. This project investigates potentially powerful precision medicine approaches to this problem by examining the impact of baseline covariates on longitudinal outcome trajectories. For example, on average, patients treated with an active drug versus placebo may have similar trajectories, but specific trajectory shapes may be unique to individuals treated with the active medication. We introduce a scalar measure from a functional observation based on an average tangent slope (ATS) to extract information of the given trajectories. Since the tangent slope represents an instantaneous rate of improvement or deterioration, an ATS can produce a useful summary from a functional observation that incorporates the shape of the trajectory. Based on the ATS, we develop methods that optimally separate longitudinal trajectories across treatment groups using the linear combinations of baseline characteristics (e.g., "biosignature"), in terms of a Kullback-Leibler distance on the distributions of outcome trajectory coefficients as well as the square difference between ATS of treatment groups.

Machine- Understandable Saliency Estimation In Natural Language Processing

♦*SHUNYAN LUO, Zhou Yang and Fang Jin*
George Washington University
shine_lsy@gwmail.gwu.edu

Deep neural networks have achieved tremendous success and even surpassed human performance in various areas. However, the usage of deep learning models is largely limitedin some tasks like automating medical diagnosis, resumescreening, etc., where interpretability matters significantly.In the neural language processing domain, attention mecha-nism is treated as inherent explanation on deep NLP models.However, there is lots of buzz on its fidelity. As a result, weadopted linearly gradient gradient(LEG), a saliency estima-tion framework origin in computer vision and proposed amodel-understandable saliency estimation(MUSE) by making normalized linear Gaussian perturbations on the encodinglayer to interpret NLP models. A variant of MUSE with regularization is proposed to enhance interpretability to humans.The interpretability is verified through a crowd-sourced human evaluation study while sensitivity analysis is conducted to validate fidelity by interpreting LSTM and BERT model on three dataset among state-of-art interpretation methods.MUSE achieves excellence performance on both studies. We also conduct sanity check to measure the sensitivity of MUSE.

LRPT: A deep learning-based dynamic framework to improve resistance prediction on longitudinal bacterial samples via Transfer Learning

♦*Yiqing Wang*[1]*, Shidan Wang*[2]*, Jiwoong Kim*[2]*, Sen Yang*[1]*, Guanghua Xiao*[2] *and Xiaowei Zhan*[2]
[1]Southern Methodist University
[2]Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America
lucy@smu.edu

Antibiotic resistance is a global health and development threat. Recently, predicting antibiotic resistance based on bacterial genetics shed new light to improve AR stewardship. In this work, we seek to achieve a highly accurate prediction model for longitudinal samples that are commonly seen in chronic infectious diseases. However, most existing algorithms do not model the emerging microbial genetic features across time, which leads to inferior performance. To overcome this limitation, we developed a deep learning-based dynamic framework, Longitudinal Resistance Prediction based on Transfer Learning (LRPT). It's designed to predict new bacterial sample's resistance, by transferring information learned from previous data of the same host. We tested it using 115 bacterial samples from five hosts, and it achieved comparatively promising prediction performance: 25% to 82% increased prediction accuracy. Compared with other commonly used machine learning algorithms, LRPT could flexibly accommodate the new genetic features discovered at later timepoints. It imposes few computational burdens. It is also applicable to longitudinal studies where predictions are needed across time. It is helpful to resolve the antibiotic resistance problem and other time-related biomedical problems.

MB-SupCon: An Integrative Modeling Framework to Improve Microbiome-based predictive models via Supervised Contrastive Learning

♦*Sen Yang*[1]*, Shidan Wang*[2]*, Yiqing Wang*[1]*, Jiwoong Kim*[2]*, Dajiang Liu*[3]*, Guanghua Xiao*[2] *and Xiaowei Zhan*[2]
[1]Department of Statistical Science, Southern Methodist University
[2]Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwestern Medical Center
[3]Department of Public Health Sciences, Pennsylvania State University
senyang@smu.edu

There are thousands of microorganisms living in human bodies. Those with distinctive abundance can be used as biomarkers as they are closely related to clinical covariates, including demographics (e.g., gender), physiological conditions (e.g., obesity) and clinical features (e.g., disease status) of the host. Thus, based on microbiome data, predicting clinical covariates and identifying influential microbes is desired and critically important. We propose a novel integrative modeling framework, MB-SupCon (Microbiome-based Supervised Contrastive Learning Framework), to address the problem. By integrating other omics data (e.g., metabolome), MB-SupCon maximizes the similarity of multi-omics information and makes better predictions based on contrastive embeddings of microbiome. Furthermore, MB-SupCon can identify the most contributing microbes to each covariate by utilizing layer-wise relevance propagation (LRP). With a well-trained model, we can make predictions on microbiome data only. For demonstration, we apply MB-SupCon to a longitudinal host-microbe dynamics study of 720 paired samples from prediabetic patients. MB-SupCon overperforms existing prediction methods (e.g., regression and mixOmics) regarding the accuracy and mean squared error. Moreover, the embedding shows more separable clusters for different covariate groups in the lower-dimensional space. With these advantages, MB-SupCon can benefit disease research and further advance the discovery of clinical relevance of microbiome and the host.

Reluctant Interaction Modeling in GLM

♦*Hai Lu and Guo Yu*
University of California, Santa Barbara

hailu@ucsb.edu

While including pairwise interactions in a regression model can better approximate the response surface, fitting such an interaction model is a well-known difficult problem. In particular, analyzing contemporary high-dimensional datasets often leads to extremely large-scale interaction modeling problems, where the challenge is posed to identify important interactions among millions or even billions of candidate interactions. While several methods have recently been proposed to tackle this challenge, they are mostly designed by (1) focusing on linear models with interactions and (or) (2) assuming the hierarchy assumption among the important interactions. In practice, however, neither of these two building blocks has to hold. In this talk, we propose an interaction modeling framework in generalized linear models (GLMs) which is free of any assumptions on hierarchy. The basic premise is a non-trivial extension of the reluctance principle to interaction selection in GLMs, where a main effect is preferred over an interaction if all else is equal. The proposed method is easy to implement, and is highly scalable to large-scale datasets. Numerical results show that the proposed method does not sacrifice any statistical performance in the presence of significant computational gain.

Sensitivity Analysis of Causal Treatment Effect Estimation for Clustered Observational Data with Unmeasured Confounding

♦*Yang Ou, Lu Tang and Chung-Chou Chang*
University of Pittsburgh
yao11@pitt.edu

Estimation of causal treatment (exposure) effect in presence of unmeasured treatment-outcome confounders often produces bias for data from an observational study. Sensitivity analysis is a useful tool for accessing how robust a treatment effect estimation with many methods developed to date. This paper proposes a new sensitivity analysis technique for normally distributed outcomes under clustered observational design with single study or multiple studies (meta-analysis) involved. Unlike the existing sensitivity analysis methods, our methods do not require additional assumptions on the number of unmeasured confounders, correlation between measured and unmeasured confounders, and interaction between measured confounders and the treatment. Our methods are easy to implement using the standard statistical software packages.

On the Bayesian Multiple Index Additive Models

♦*Zhengkang Liang and Zhigen Zhao*
Temple University
tuh38007@temple.edu

We consider the Bayesian multi-index additive model (BMIAM): $Y_i = \sum_{d=1}^{D} f_d\left(X_i'\beta_d\right) + \epsilon_i$. The index is parameterized by polar coordinates and the function of each additive component is approximated by the Bayesian B-splines. The number of directions is selected by a modified BIC approach. We developed an MCMC algorithm. It has been shown through both simulation and real data analysis that the proposed method works substantially better than existing methods, such as MAVE, random forest, and others.

Mixture of Shape-on-Scalar Regression Models: Going Beyond Prealigned Non-Euclidean Responses

♦*Yuanyao Tan*[1] *and Chao Huang*[2]
[1]Florida State University
[2]chuang7@fsu.edu
yt20ba@my.fsu.edu

Due to the wide applications of shape data analysis in medical imaging, computer vision, and many other fields, it is of great interest to cluster objects and recover the underlying sub-group structure according to their shapes and covariates in Euclidean space (e.g., age and diagnostic status). However, this clustering task faces four challenges including (i) non-Euclidean space, (ii) misalignment of shapes due to pre-processing steps and imaging heterogeneity, (iii) complex spatial correlation structure, and (iv) geodesic variation associated with some covariates. In order to address these challenges, we propose a mixture of geodesic factor regression model (M-GeoFARM). In each cluster, a geodesic regression structure including covariates of interest and alignment step is established along with the Riemannian Gaussian distribution in the pre-shape space, and a latent factor model is built in the tangent space. In addition, a Monte Carlo EM algorithm is provided for the parameter estimation procedure. Finally, both simulation studies and real data analysis are conducted to compare the clustering performance of M-GeoFARM with other existing methods.

An Eigenmodel for Dynamic Multilayer Networks

♦*Joshua Loyal and Yuguo Chen*
University of Illinois at Urbana-Champaign
jloyal2@illinois.edu

Dynamic multilayer networks frequently represent the structure of multiple co-evolving relations; however, statistical models are not well-developed for this prevalent network type. Here, we propose a new latent space model for dynamic multilayer networks. The key feature of our model is its ability to identify common time-varying structures shared by all layers while also accounting for layer-wise variation and degree heterogeneity. We establish the identifiability of the model's parameters and develop a structured mean-field variational inference approach to estimate the model's posterior, which scales to networks previously intractable to dynamic latent space models. We demonstrate the estimation procedure's accuracy and scalability on simulated networks. We apply the model to two real-world problems: discerning regional conflicts in a data set of international relations and quantifying infectious disease spread throughout a school based on the student's daily contact patterns.

A Sparse Projection Regression Framework for Generalized Eigenvalue Problems

♦*Dylan Molho*[1]*, Yuying Xie*[1] *and Qiang Sun*[2]
[1]Michigan State University
[2]University of Toronto
molhody1@msu.edu

This paper proposes a unified sparse projection regression framework for estimating generalized eigenvector problems. Unlike existing work, we reformulate the sequence of computationally intractable non-convex generalized Rayleigh quotient optimization problems into a computationally efficient simultaneous linear regression problem, padded with a sparse penalty to deal with high dimensional predictors. Theoretically, we show the reformulated linear regression problem is able to recover the same projection space obtained by the original generalized eigenvalue problem. Statistically, we establish the nonasymptotic error bound for the proposed estimator. We showcase the applications of our method by considering three iconic problems in statistics: the sliced inverse regression, the linear discriminant analysis, and the canonical correlation analysis. Numerical studies and real data applications lend strong support to our proposed methodology.

Statistical Learning in Preclinical Drug Proarrhythmic Assessment

♦*Nan Miles Xi*[1]*, Yu-Yi Hsu*[2]*, Qianyu Dang*[2] *and Dalong Patrick Huang*[2]
[1]Loyola University Chicago

[2]FDA
mxi1@luc.edu

Torsades de pointes (TdP) is an irregular heart rhythm characterized by faster beat rates and potentially could lead to sudden cardiac death. Much effort has been invested in understanding the drug-induced TdP in preclinical studies. However, a comprehensive statistical learning framework that can accurately predict the drug-induced TdP risk from preclinical data is still lacking. We proposed ordinal logistic regression and ordinal random forest models to predict low-, intermediate-, and high-risk drugs based on datasets generated from two experimental protocols. Leave-one-drug-out cross-validation, stratified bootstrap, and permutation predictor importance were applied to estimate and interpret the model performance under uncertainty. The potential outlier drugs identified by our models are consistent with their descriptions in the literature. Our method is accurate, interpretable, and thus useable as supplemental evidence in the drug safety assessment.

Artificial intelligence-driven radiogenomic analysis framework with mediation analysis for identifying prognostic radiogenomic biomarkers in breast cancer

♦*Qian Liu and Pingzhao Hu*
University of Manitoba
qianl@myumanitoba.ca

Background: Radiogenomics is to study medical images and genomic profiles jointly for critical clinical problems. We proposed a novel framework to identify breast cancer (BC) prognostic radiogenomic biomarkers. Methods: Bayesian tensor factorization was used to extract multi-omics features from gene expression, DNA methylation, and copy number variation data of 762 BC patients. An explainable deep learning (DL) model was built to extract multimodal MRI features for 61 of these BC patients. LASSO models were trained to impute the MRI features whose prognostic significance was then estimated. Mediation analyses were performed to explore the biological mechanisms of the identified biomarkers. Traditional semi-auto radiomic features and well-known gene expression features acted as baselines. Results: Three DL-based multi-modal MRI radiogenomic biomarkers were successfully identified, which have significant differences in overall survival (log-rank test, Bonferroni corrected p-value¡0.05). The most significant one was associated with 10 BC risk genes (such as APITD1, HNF4) and several metabolism-related pathways (Purine metabolism pathway and Tryptophan metabolism pathway), which has a significant mediation effect on the relationship between the function of natural killer cells and the BC survival (adjusted p-value¡0.002). Conclusion: The results may promote MRI as a non-invasive examination of probing BC prognosis and multi-level molecular status.

A Novel Model Checking Approach for Dose-Response Relationships

♦*Yu Xia*[1]*, Xinmin Li*[2]*, Hua Liang*[1] *and Shunyao Wu*[2]
[1]George Washington University
[2]Qingdao University
xiayu@gwu.edu

We propose a test for assessing nonlinear dose-response models based on a Cramer-von Mises statistic. We establish the asymptotic distribution of the test and demonstrate that the test can detect the local alternative converging to the null at the parametric rate $\frac{1}{\sqrt{n}}$. We provide a bootstrap resampling technique to calculate the critical values. It is observed that the test has good power performance in small sample sizes. We apply the proposed method to analyze 250 datasets from a pharmacologic study, and conduct two small simulation experiments to explore the numerical performance of the proposed test and compare one commonly used test in practice.

Semiparametric marginal regression analysis of clustered multistate process data

♦*Wenxian Zhou*[1]*, Giorgos Bakoyannis*[1]*, Ying Zhang*[2] *and Constantin T. Yiannoutsos*[1]
[1]Indiana University
[2]University of Nebraska Medical Center
wz11@iu.edu

Clustered multistate process data are commonly encountered in multicenter observational studies and clinical trials. However, there are no methods for semiparametric regression on state occupation probabilities with such clustered multistate data. In this work, we address this issue by proposing a framework for semiparametric marginal regression analysis of state occupation probabilities for clustered multistate processes. The estimation procedure involves solving a set of functional generalized estimating equations. The proposed method does not impose Markov assumptions or assumptions regarding the structure of the within-cluster dependence, and allows for informative cluster size. The asymptotic properties of the proposed estimators for functional regression coefficients parameters are rigorously established using empirical process theory and a non-parametric hypothesis testing procedure of covariate effects is provided. Simulation studies show that the proposed method performs well in finite samples and that ignoring the within-cluster dependence and informative cluster size leads to invalid inferences. The proposed method is illustrated using clustered multistate data from a multicenter randomized clinical trial on recurrent or metastatic squamous-cell carcinoma of the head and neck.

A Bayesian Approach to Variable Selection in Binary Quantile Regression

♦*Mai Dao*[1]*, Min Wang*[2] *and Souparno Ghosh*[3]
[1]Texas Tech University
[2]University of Texas at San Antonio
[3]University of Nebraska-Lincoln
mai.dao@ttu.edu

In this talk, we present a Bayesian hierarchical modeling framework for simultaneously conducting parameter estimation and variable selection in binary quantile regression. We first impose the asymmetric Laplace distribution on the model errors and then specify a quantile-dependent prior for the regression coefficients, which allows researchers to set different priors for modeling different orders of quantiles and thus yields great flexibility in Bayesian quantile modeling. By utilizing the normal-exponential mixture representation of the asymmetric Laplace distribution, we propose a novel three-stage computational scheme starting with an expectation-maximization algorithm and then the Gibbs sampler followed by an importance re-weighting step to draw independent Markov chain Monte Carlo samples from the full conditional posterior distributions of the unknown parameters. Simulation studies are conducted to compare the performance of the proposed Bayesian method with that of several existing ones in the literature. Finally, real-data applications are provided for illustrative purposes.

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

♦*Siddhesh Kulkarni, Jeremy Gaskins and Subhadip Pal*
University of Louisville
siddhesh.kulkarni@louisville.edu

In principle, it can be challenging to integrate data measured on same individuals occurring from different experiments and model

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

Statistical analyses of housekeeping genes' methylation patterns in breast cancer

Shuying Sun
Texas State University
ssun5211@yahoo.com

DNA methylation is an epigenetic event. It occurs when a methyl group is added to a cytosine base that is paired with a guanine (i.e., a CG site). DNA methylation plays an important role in regulating gene expression. Many tumor suppressor genes are found to be methylated and associated with different types of cancers including breast cancer. However, few studies are reported on the methylation patterns of housekeeping genes, which maintain the basic functions of cells. In this study, we will present our statistical analysis results on housekeeping genes' methylation patterns in breast cancer patients using publicly available data. In particular, we analyze the methylation patterns by comparing breast tumors with adjacent normal tissues. We will also compare the living patients with the dead samples to identify the housekeeping genes with stable methylation (SM) as well as the ones with differential methylation (DM). We will also show genetic pathway and network analyses of these different SM and DM sets of genes. Our findings will provide researchers with a new and improved understanding of housekeeping genes' methylation patterns in breast cancer.

Application of machine learning methods in clinical trial design with response?adaptive randomization

♦*Yizhuo Wang*[1]*, Xuelin Huang*[1]*, Ziyi Li*[1] *and Bing Carter*[2]
[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center
[2]Department of Leukemia, The University of Texas MD Anderson Cancer Center
ywang70@mdanderson.org

A key component for the success of response-adaptive randomization is a good prediction algorithm for patients' response to treatments. Machine learning methods have successfully demonstrated their superior prediction performance in many applications, but have not been applied in adaptive clinical trials. In this article, we incorporate nine machine learning methods, such as gradient boosting machine, random forest and artificial neural network, in clinical trial design. Realizing that not a single method may fit all trials well, we also use an ensemble of these nine methods. We evaluate their performance through the benefits for individual trial participants, such as the percentage of patients who receive their optimal treatments and individual loss. To avoid the potential bias introduced by the adaptive scheme, we use inverse probability of treatment weighted method to estimate the average treatment effect and power. Simulation studies show that use of machine learning methods results in more personalized treatment assignment and higher overall response rates among trial participants. Compared with the nine individual methods of machine learning, the ensemble approach achieves the highest response rate and assigns the largest percentage of patients to their optimal treatments. The proposed methods are applied to a real-world leukemia study.

Comparison of data-driven subgroup identification methods for binary and time-to-event outcome

♦*Yujie Zhao*[1]*, Ahrim Youn*[2]*, Yanping Chen*[2] *and Casey Xu*[2]
[1]The University of Texas, MD Anderson Cancer Center
[2]Bristol Myers Squibb
yujie.zhao@uth.tmc.edu

In clinical trials, patients with different baseline characteristics may show heterogeneous treatment effect. Identification of patient subgroups who are more likely to respond to a treatment than others plays an essential role to increase the efficiency of drug development and support precision medicine. Despite the increasing interest in predictive enrichment strategies in trial design, there is no comprehensive comparative study to evaluate the performance of data-driven subgroup identification methods, especially for both binary outcome and time-to-event outcome. In this project, we review 6 methods developed in statistics and machine learning community with publicly available R packages. We evaluate the operating characteristics of these models using simulated randomized trial data of various size based on four major criteria: (1) power to detect predictive biomarkers and associated cutoffs (2) rate of false biomarker identification (3) sensitivity and specificity to select the patient subgroups who benefit from the treatment (4) bias of treatment effect estimation in selected subgroup.

Robust Inference for Linear Models under Huber's Contamination Model

♦*Peiliang Zhang*[1]*, Wen-Xin Zhou*[2] *and Zhao Ren*[1]
[1]University of Pittsburgh
[2]University of California, San Diego
PEZ35@pitt.edu

We study the robust estimation and inference problem for linear models in the increasing dimension regime. Given random design, we consider the conditional distributions of error terms are contaminated by some arbitrary distribution (possibly depending on the covariates) with proportion p but otherwise can also be heavy-tailed and asymmetric. Under the setting of Huber's contamination model, we prove that simple robust M-estimators (like Huber or smoothed Huber), with an additional intercept added in the model, can achieve the optimal minimax rate of convergence under the l2 loss. In addition, two types of confidence intervals with root-n consistency are provided by a multiplier bootstrap technique when the necessary condition on contamination proportion $p = o(1/n^1/2)$ holds. For a larger p, we further propose a debiasing procedure to reduce the potential bias caused by contamination, and prove the validity of the debiased confidence interval. Our method can be extended to the communication-efficient distributed estimation and inference setting in a straightforward way. A comprehensive simulation study exhibits the effectiveness of our proposed inference procedures.

Application of multiple criteria decision analysis (MCDA) in early phase drug development

♦*Weibin Zhong*[1]*, Alan Wu*[2]*, Casey Xu*[2]*, Ahrim Youn*[2] *and Jun Zhang*[2]
[1]Bristol Myers Squibb and George Mason University

[2]Bristol Myers Squibb
Weibin.Zhong@bms.com

The multiple criteria decision analysis (MCDA) methods have been widely applied to support decision makers in evaluating alternatives by considering multiple criteria. It can be used to capture and quantify benefits, risks, and uncertainties of alternatives to aid the decision-making process by considering an explicit set of criteria. The MCDA method is recommended for drug benefit-risk assessment and support treatment selection in early phase development. This presentation reviews MCDA methods with different weighting methods. We compare these methods by evaluating the potential advantages and pitfalls. R shiny apps of the MCDA methods with different weighting methods are developed to help make comparisons of the alternatives. Real applications for some drug benefit-risk data are conducted by using the MCDA R shiny apps.

Applying Group Sequential Design with Multiple Comparison Consideration in Confirmatory Clinical Trial

♦*Shuyu Chu and Min Chen*
BMS
shuyu.chu@bms.com

The methodology of planning, conducting, and analyzing group sequential design in clinical trials with one primary hypothesis is well developed. Statistical methodology that addresses the multiplicity issue is also advanced and widely applied. However, as confirmatory clinical trials become increasingly complex and involve testing multiple hypotheses group sequentially, the multiplicity issue then expands in two dimensions: multiple time points and multiple hypotheses. It becomes challenging to strongly control the familywise error rate (FWER) at a pre-specified significance level for each individual hypothesis as well as across all hypotheses. We will present a specific clinical design issue that may occur in designing a late phase oncology study. Two possible testing designs and strategies, including hierarchical testing and recently developed graphical approaches together with group sequential analysis will be discussed. We will also illustrate and compare those methods through numerical studies.

An Application of Predictive Modelling in Supporting Fedratinib Regulatory Filing

♦*Jun Zhang, Alan Wu and Ivan Chan*
BMS
samulezj@gmail.com

Regulatory decision on drug submission is evidence-based. In some circumstance, key trial results may be limited due to unexpected events, such as termination by sponsor or clinical hold. Those events may subsequently impact the maturity of the study data that includes primary endpoints, long-term efficacy and safety data which are key elements in regulator's risk-benefit evaluations. Without successfully addressing the impacts of study termination on those key data, the application is at risk. In these circumstances, creative tools such as Predictive modelling could be used. In this abstract we have presented, a predictive modelling that was used to support a regulatory submission on myelofibrosis (MF) indication. In MF patients, spleen volume reduction (SVR) is the primary efficacy measure. SVR data, in 2L MF study was impacted due to clinical hold. A parametric predictive model was built based on longitudinal SVR data to bridge 1L and 2L MF studies. The model not only strengthened the primary endpoint in 2L MF study but also predicted a higher SVR rate and long-term clinical benefits. This modeling work helped the regulators to assess the major concern of truncated SVR data and led to the regulatory approval in 1L and 2L MF.

New Class of Distortion Risk Measures and Their Estimation

Xiwen Wang
Central Michigan University
wang18x@cmich.edu

In this paper, we present a new method to construct new classes of distortion functions. A distortion function maps the unit interval to the unit interval and has characteristics of a cumulative distribution function. The method is based on the transformation of an existing non-negative random variable whose distribution function, named the generating distribution, may contain more than one parameter. The coherency of the resulting risk measure is ensured by restricting the parameter space on which the distortion function is concave. We study the cases when the generating distributions are exponentiated exponential and Gompertz distributions. Closed-form expressions for risk measures are derived for uniform, exponential, and Lomax losses. Numerical and graphical results are presented to examine the effects of parameter values on the risk measures. We then propose a simple estimator of risk measures and conduct simulation studies to compare and demonstrate the performance of the proposed estimator for various losses.

On Construction and Estimation of Stationary Mixture Transition Distribution Models

♦*Xiaotian Zheng, Athanasios Kottas and Bruno Sansó*
University of California Santa Cruz
xzheng24@ucsc.edu

Mixture transition distribution time series models build high-order dependence through a weighted combination of first-order transition densities for each one of a specified number of lags. We present a framework to construct stationary mixture transition distribution models that extend beyond linear, Gaussian dynamics. We study conditions for first-order strict stationarity which allow for different constructions with either continuous or discrete families for the first-order transition densities given a pre-specified family for the marginal density, and with general forms for the resulting conditional expectations. Inference and prediction are developed under the Bayesian framework with particular emphasis on flexible, structured priors for the mixture weights. Model properties are investigated both analytically and through synthetic data examples. Finally, Poisson and Lomax examples are illustrated through real data applications.

Session 54: Nonconvex Optimization in Statistics

Linear Polytree Structural Equation Models: Structural Learning and Inverse Correlation Estimation

Xingmei Lou[1]*, Yu Hu*[2] *and* ♦*Xiaodong Li*[1]
[1]UC Davis
[2]Hong Kong University of Science and Technology
xysddlxd@gmail.com

We are interested in the problem of learning the directed acyclic graph (DAG) when data are generated from a linear structural equation model (SEM) and the causal structure can be characterized by a polytree. Specially, under both Gaussian and sub-Gaussian models, we study the sample size conditions for the well-known Chow-Liu algorithm to exactly recover the equivalence class of the polytree, which is uniquely represented by a CPDAG. We also study the error rate for the estimation of the inverse correlation matrix under such models. Our theoretical findings are illustrated by comprehensive numerical simulations, and experiments on benchmark data also demonstrate the robustness of the method when the ground truth graphical structure can only be approximated by a polytree.

Compositional Optimization under Misspecification

Ethan Fang
Penn State
xxf13@psu.edu

As systems grow in size, scale, and intricacy, the challenges of misspecification become even more pronounced. In this paper, we focus on parametric misspecification in regimes complicated by risk and nonconvexity. When this misspecification may be resolved via a parallel learning process, we develop data-driven schemes for resolving a broad class of misspecified stochastic compositional optimization problems. Notably, this rather broad class of compositional problems can contend with challenges posed by diverse forms of risk, dynamics, and nonconvexity, significantly extending the reach of such avenues. Specifically, we consider the minimization of a stochastic compositional function $[f([g(x;\theta_2^*,\xi_2)];\theta_1^*,\xi_1)]$ over a closed and convex set in a regime where parameters θ_1^* and θ_2^* are unknown, and ξ_1 and ξ_2 are two suitably defined random variables. Existing algorithms can accommodate settings where parameters θ_1^* and θ_2^* are known, but efficient first-order schemes are hitherto unavailable for the imperfect information compositional counterparts. Via a data-driven compositional optimization (DDCO) approach, we develop asymptotic and rate guarantees for unaccelerated and accelerated schemes for convex, strongly convex, and nonconvex problems in a two-level regime. Additionally, we extend the accelerated schemes to the general T-level setting. Notably, the non-asymptotic rate guarantees in all instances show no degradation from the rate statements obtained in a correctly specified regime. Our numerical experiments support the theoretical findings based on the resolution of a three-level compositional risk-averse optimization problem.

Sample complexity of Q-learning: sharper analysis and variance reduction

Gen Li[1], ♦*Yuting Wei*[2], *Yuejie Chi*[3], *Yuantao Gu*[4] *and Yuxin Chen*[1]
[1]Princeton University
[2]University of Pennsylvania
[3]CMU
[4]Tsinghua University
ytwei@wharton.upenn.edu

Asynchronous Q-learning aims to learn the optimal action-value function (or Q-function) of a Markov decision process (MDP), based on a single trajectory of Markovian samples induced by a behavior policy. Focusing on a γ-discounted MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$, we demonstrate that the ℓ_∞-based sample complexity of classical asynchronous Q-learning – namely, the number of samples needed to yield an entrywise ϵ-accurate estimate of the Q-function – is at most on the order of $\frac{1}{\mu_{\min}(1-\gamma)^5\epsilon^2}+\frac{t_{mix}}{\mu_{\min}(1-\gamma)}$ up to some logarithmic factor, provided that a proper constant learning rate is adopted. Here, t_{mix} and $\mu_{\min}$ denote respectively the mixing time and the minimum state-action occupancy probability of the sample trajectory. The first term of this bound matches the complexity in the case with independent samples drawn from the stationary distribution of the trajectory. The second term reflects the expense taken for the empirical distribution of the Markovian trajectory to reach a steady state, which is incurred at the very beginning and becomes amortized as the algorithm runs. Encouragingly, the above bound improves upon the state-of-the-art result by a factor of at least $|\mathcal{S}||\mathcal{A}|$. Further, the scaling on the discount complexity can be improved by means of variance reduction.

Asymptotic Analysis of Accelerated Stochastic Gradient Descent

Shang Wu
Fudan University
shangwu@fudan.edu.cn

Optimization takes an important role in various fields. First-order methods like gradient descent and its variants exhibit good performance in practice. Nesterov presents a new scheme to accelerate the vanilla gradient descent. He employs the auxiliary variables and shows the improvement of convergence rate from $O(k^{-1})$ to $O(k^{-2})$, where k denotes the number of algorithm iterations. One can connect the algorithms with ordinary differential equations. For Nesterov's accelerated gradient descent, I prove that the difference between the algorithm sequence and the solution to the corresponding ordinary differential equation is uniformly bounded by $O(\delta^{(1/2)}|\log(\delta)|)$ over a fixed interval, where delta is the step size. In machine learning, stochastic gradient descent is a useful and efficient algorithm for solving optimization problems. When the objective function can only be estimated by the average of loss functions based on a large amount of data, it's very time-consuming to calculate the gradient using all data at every step. In stochastic gradient descent, it updates the variable using the gradient of the average loss function based on the random mini-batch samples. I apply Nesterov's acceleration method on stochastic gradient descent and study the random sequence generated by the algorithm. We connect the stochastic algorithms with the second-order stochastic differential equations. I analyze the difference between the random sequence and the solution to the corresponding SDE. As a result, when the step size tends to zero and the mini-batch sample size goes to infinity, I establish the asymptotic theory for the algorithms.

Session 55: Transforming Industries with Advanced Data Science Solutions

Lessons learned from AlphaGo Zero to AlphaFold 2

Haoda Fu
Eli Lilly and Company
NA

We reviewed the two most impactful AI algorithms as AlphaGo Zero and AlphaFold 2, and discuss how it will impact drug discovery, development, and commercialization.

A few things about product data scientist

♦*Jie Cheng and Yuan Jin*
Google
NA

An overview on what product data scientist do and a few things we learned on how to be an effective product data scientist.

A Time To Event Framework For Multi-touch Attribution

Dinah Shender[1], *Ali Nasiri Amini*[1], *Xinlong Bao*[1], *Mert Dikmen*[1], *Amy Richardson*[2] *and* ♦*Jing Wang*[1]
[1]Google
[2]At Google when this work was done
jiwang@google.com

One of the promises of online advertising has been the ability to tie together ad views or clicks with actual outcomes (e.g. purchases, website visits, etc), also known as conversions, in order to give advertisers more insight into the effectiveness of their ads. In particular, advertisers are interested in understanding the relative contributions of the multiple ads a user may see prior to an observed conversion, a problem known as multi-touch attribution (MTA), in order to adjust their budget and bidding decisions accordingly. This presents

a challenging problem due to the incomplete (or streaming) nature of the data, as well as the need to allow an ad's effect to change over time. We describe an MTA system, consisting of a model for user conversion behavior and a credit assignment algorithm, that satisfies these requirements. Our model for user conversion behavior treats conversions as occurrences in an inhomogeneous Poisson process, while our attribution algorithm is based on iteratively removing the last ad in the path.

Session 56: Statistical Considerations for Data Integration and Data Privacy

Integrating Information from Existing Risk Prediction Models with No Model Details

♦*Peisong Han, Jeremy Taylor and Bhramar Mukherjee*
University of Michigan
peisong@umich.edu

Consider the setting where (i) individual-level data are collected to build a regression model for the association between an event of interest and certain covariates, and (ii) some risk calculators predicting the risk of the event using less detailed covariates are available, possibly as algorithmic black boxes with little information available about how they were built. We propose a general empirical-likelihood-based framework to integrate the rich auxiliary information contained in the calculators into fitting the regression model in order to improve the efficiency for the estimation of regression parameters. As an application, we study the dependence of the risk of high grade prostate cancer on both conventional risk factors and newly identified molecular biomarkers by integrating information from the Prostate Biopsy Collaborative Group (PBCG) risk calculator.

Privacy-protecting Cox Proportional Hazards Regression for Distributed Data Networks Using Summary-level Information

Dongdong Li[1], *Wenbin Lu*[2], *Di Shu*[3], *Sengwee Toh*[1] *and* ♦*Rui Wang*[1]
[1]Harvard Pilgrim Health Care Institute
[2]North Carolina State University
[3]University of Pennsylvania
rwang@hsph.harvard.edu

Individual-level data sharing across multiple sites can be infeasible due to privacy and logistical concerns. This article proposes a general methodology to fit Cox proportional hazards models without sharing individual-level data. 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, and can 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 step of 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, the alternative distribution regression, 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.

Gaussian Differential Privacy

Weijie Su
University of Pennsylvania
suw@wharton.upenn.edu

Privacy-preserving data analysis has been put on a firm mathematical foundation since the introduction of differential privacy (DP) in 2006. This privacy definition, however, has some well-known weaknesses: notably, it does not tightly handle composition. In this talk, we propose a relaxation of DP that we term "f-DP", which has a number of appealing properties and avoids some of the difficulties associated with prior relaxations. First, f-DP preserves the hypothesis testing interpretation of differential privacy, which makes its guarantees easily interpretable. It allows for lossless reasoning about composition and post-processing, and notably, a direct way to analyze privacy amplification by subsampling. We define a canonical single-parameter family of definitions within our class that is termed "Gaussian Differential Privacy", based on hypothesis testing of two shifted normal distributions. We prove that this family is focal to f-DP by introducing a central limit theorem, which shows that the privacy guarantees of any hypothesis-testing based definition of privacy (including differential privacy) converge to Gaussian differential privacy in the limit under composition. This central limit theorem also gives a tractable analysis tool. We demonstrate the use of the tools we develop by giving an improved analysis of the privacy guarantees of noisy stochastic gradient descent. This is joint work with Jinshuo Dong and Aaron Roth.

Distributed algorithms for mixed effects models

♦*Yong Chen and Chongliang Luo*[1]
[1]Washington University at St Louis
NA

Linear mixed models (LMMs) are commonly used in many areas including epidemiology for analyzing multi-site data with heterogeneous site-specific random effects. However, due to the regulation of protecting patients' privacy, sensitive individual patient data (IPD) are usually not allowed to be shared across sites. In this paper we propose a novel algorithm for distributed linear mixed models (DLMMs). Our proposed DLMM algorithm can achieve exactly the same results as if we had pooled IPD from all sites, hence the lossless property. The DLMM algorithm requires each site to contribute some aggregated data (AD) in only one iteration. We apply the proposed DLMM algorithm to analyze the association of length of stay of COVID-19 hospitalization with demographic and clinical characteristics using the administrative claims database from the UnitedHealth Group Clinical Research Database.

Session 57: Recent developments in survival and recurrent event data analysis

Testing the proportional hazard assumption under monotonicity constraints

Huan Chen and ♦*Chuan-Fa Tang*
University of Texas at Dallas
chuan-fa.tang@utdallas.edu

Cox proportional hazard model has been proposed for decades, however, the proportional hazard assumption can be hard to check in practice. In many applications, instead of proportionality, it is reasonable to assume the monotonicity of the hazard function and many nonparametric estimators has been proposed, such as Groeneboom and Wellner (1992), Banerjee (2007), and Chung et al. (2018). We study the estimators and propose a test for the proportional hazard assumption under monotonicity constraints. Our simulation results show that the proposed test has well-controlled size and consistency.

Weighted least-squares estimation for semiparametric accelerated failure time model with regularization

Ying Chen, Chuan-Fa Tang, ♦Sy Han Chiou and Min Chen
University of Texas at Dallas
schiou@utdallas.edu

Clustered failure times often arise from case-cohort design when the outcome is rare and the main covariates are expensive to measure. A proper statistical model needs to correct the sampling bias and account for within-cluster dependence. Accelerated failure time models for case-cohort data have been developed but most are based on a rank-based estimating approach which is less intuitive than the least-squares approach. In this paper, we propose a weighted least-squares procedure incorporating inverse prob- ability weights for regression parameters estimation. The proposed method is also extended to multivariate case-cohort data accounting for within-cluster dependency using a generalized estimating equation. Large-scale simulation results show our proposed method can provide good point estimators for both univariate and multivariate cases. Higher estimation efficiency can be achieved when the within-cluster dependency is taken into account. An application to retrospective dental study demonstrates the utility of the proposed method in routine data analysis.

Estimation and Model Checking for General Semiparametric Recurrent Event Models with Informative Censoring

Hung-Chi Ho
China Medical University and Hospital, Taiwan
hermitianho@gmail.com

This research investigates novel semiparametric intensity models with a multiplicative frailty for recurrent event data in the presence of informative censoring. The proposed estimation methods require neither a specification of the link function in the model nor a distributional assumption on the unobserved frailty. Specifically, estimation procedures are tailored for situations where the joint conditional distribution of the recurrent event times is non-degenerate in some covariates or degenerate in all covariates. A systematic model selection procedure is further developed to identify the functional form of the rate function, examine the correctness of a single-index model formulation, and select the link function of a single-index rate model. In addition, large-sample properties of the estimators and model selection statistics are established under some regularity conditions. Extensive simulation studies and analyses of two data examples are also presented to illustrate the proposed methodology.

Unification of semicompeting risks analysis through causal mediation modeling

♦*Jin-Chang Yu and Yen-Tsung Huang*
Institute of Statistical Science, Academia Sinica, Taipei, Taiwan
yujihchang@stat.sinica.edu.tw

Semicompeting risks represent a common problem where an intermediate event and a terminal event are both of interest, but only the former may be truncated by the latter. Copula, frailty and multistate models serve as well-established analytics for semicompeting risks. Here, we cast the semicompeting risks in a causal mediation framework, with the intermediate event as a mediator and the terminal event as an outcome. We define the indirect and direct effects as the effects of an exposure on the terminal event mediated and not mediated through the intermediate event, respectively. We derive respective expressions of estimands for causal mediation under the copula, frailty and multistate models. Next, we propose estimators based on nonparametric maximum likelihood or U-statistics and establish their asymptotic results. Numerical studies demonstrate that the efficiency of copula models leads to potential bias due to model misspecification. Moreover, the robustness of frailty models is accompanied by a loss in efficiency, and multistate models balance the efficiency and robustness. We observe a similar feature when applying the proposed methods to a hepatitis study, indicating that hepatitis B affects mortality by increasing liver cancer incidence. Thus, causal mediation modeling provides a unified framework that accommodates various semicompeting risks models.

Session 58: Recent developments in statistical network data analysis

Community Detection on Mixture Multi-layer Networks via Regularized Tensor Decomposition

Bing-Yi JING[1], Ting LI[2], Zhongyuan LYU[1] and ♦Dong XIA[1]
[1]HKUST
[2]HK PolyU
madxia@ust.hk

We study the problem of community detection in multi-layer networks, where pairs of nodes can be related in multiple modalities. We introduce a general framework, i.e., mixture multi-layer stochastic block model (MMSBM), which includes many earlier models as special cases. We propose a tensor-based algorithm (TWIST) to reveal both global/local memberships of nodes, and memberships of layers. We show that the TWIST procedure can accurately detect the communities with small misclassification error as the number of nodes and/or the number of layers increases. Numerical studies confirm our theoretical findings. To our best knowledge, this is the first systematic study on the mixture multi-layer networks using tensor decomposition. The method is applied to two real datasets: worldwide trading networks and malaria parasite genes networks, yielding new and interesting findings.

Network Structure Inference from Grouped Observations

♦*Yunpeng Zhao[1], Peter Bickel[2] and Charles Weko[3]*
[1]Arizona State University
[2]University of California, Berkeley
[3]US Army
Yunpeng.Zhao@asu.edu

Statistical network analysis typically deals with inference concerning various parameters of an observed network. In several applications, especially those from social sciences, behavioral information concerning groups of subjects are observed. Over the past century a number of descriptive statistics have been developed to infer network structure from such data. However, these measures lack a generating mechanism that links the inferred network structure to the observed groups. In this talk, we present a model-based approach called the hub model, which belongs to a family of Bernoulli mixture models. We further present theoretical results on model identifiability, a notoriously difficult problem in Bernoulli mixture models, and estimation consistency.

Two-sample inference for network moments

♦*Yuan Zhang[1], Dong Xia[2], Qiong Wu[3] and Shuo Chen[3]*
[1]Ohio State University
[2]Hong Kong University of Science and Technology
[3]University of Maryland
yzhanghf@stat.osu.edu

In this paper, we present a novel fast and higher-order accurate two-sample inference procedure for network comparison. Our work addresses the major computational challenge in network comparison. Compared to graph-matching-based network comparison methods, our method is much more scalable and all needed components in our

test can be computed offline within each network. This provides an efficient hashing framework for maintaining and fast querying huge network database, because the main numerical features of each network can be succinctly captured by a few scalar summary statistics.

Autoregressive Networks

♦*Binyan Jiang*[1]*, Jialiang Li*[2] *and Qiwei Yao*[3]
[1]The Hong Kong Polytechnic University
[2]National University of Singapore
[3]London School of Economics
by.jiang@polyu.edu.hk

We propose a first-order autoregressive model for dynamic network processes in which edges change over time while nodes remain unchanged. The model depicts the dynamic changes explicitly. It also facilitates simple and efficient statistical inference such as the maximum likelihood estimators which are proved to be (uniformly) consistent and asymptotically normal. The model diagnostic checking can be carried out easily using a permutation test. The proposed model can apply to any network processes with various underlying structures but with independent edges. As an illustration, an autoregressive stochastic block model has been investigated in depth, which characterizes the latent communities by the transition probabilities over time. This leads to a more effective spectral clustering algorithm for identifying the latent communities. Inference for a change point is incorporated into the autoregressive stochastic block model to cater for possible structure changes. The developed asymptotic theory as well as the simulation study affirms the performance of the proposed methods. Application with three real data sets illustrates both relevance and usefulness of the proposed models.

Session 59: Recent development on spatial statistics

Bayesian space-time gap filling for inference on extreme hotspots: an application to Red Sea surface temperatures

♦*Daniela Castro-Camilo*[1]*, Linda Mhalla*[2] *and Thomas Opitz*[3]
[1]University of Glasgow
[2]HEC Lausanne
[3]Biostatistics and Spatial Processes, INRAE, Avignon, France
daniela.castrocamilo@glasgow.ac.uk

We develop a method for probabilistic prediction of extreme value hot-spots in a spatio-temporal framework, tailored to big datasets containing important gaps. In this setting, direct calculation of summaries from data, such as the minimum over a space-time domain, is not possible. To obtain predictive distributions for such cluster summaries, we propose a two-step approach. We first model marginal distributions with a focus on accurate modeling of the right tail and then, after transforming the data to a standard Gaussian scale, we estimate a Gaussian space-time dependence model defined locally in the time domain for the space-time subregions where we want to predict. In the first step, we detrend the mean and standard deviation of the data and fit a spatially resolved generalized Pareto distribution to apply a correction of the upper tail. To ensure spatial smoothness of the estimated trends, we either pool data using nearest-neighbor techniques, or apply generalized additive regression modeling. To cope with high space-time resolution of data, the local Gaussian models use a Markov representation of the Matérn correlation function based on the stochastic partial differential equations (SPDE) approach. In the second step, they are fitted in a Bayesian framework through the integrated nested Laplace approximation implemented in R-INLA. Finally, posterior samples are generated to provide statistical inferences through Monte-Carlo estimation. Motivated by the 2019 Extreme Value Analysis data challenge, we illustrate our approach to predict the distribution of local space-time minima in anomalies of Red Sea surface temperatures, using a gridded dataset (11,315 days, 16,703 pixels) with artificially generated gaps. In particular, we show the improved performance of our two-step approach over a purely Gaussian model without tail transformations.

Global Wind Modeling with Transformed Gaussian Processes

Jaehong Jeong

Hanyang University
jaehongjeong@hanyang.ac.kr

Uncertainty quantification of wind energy potential from climate models can be limited because it requires considerable computational resources and is time-consuming. We propose a stochastic generator that aims at reproducing the data-generating mechanism of climate ensembles for global annual, monthly, and daily wind data. Inferences based on a multi-step conditional likelihood approach are achieved by balancing memory storage and distributed computation for a large data set. In the end, we discuss a general framework for modeling non-Gaussian multivariate stochastic processes by transforming underlying multivariate Gaussian processes.

Modeling Spatial Data with Cauchy Convolution Processes

♦*Pavel Krupskiy*[1] *and Raphael Huser*[2]

[1]University of Melbourne

[2]King Abdullah University of Science and Technology
pavel.krupskiy@unimelb.edu.au

We study the class of models for spatial data obtained from Cauchy convolution processes based on different types of kernel functions. We show that the resulting spatial processes have some appealing tail dependence properties. We derive the extreme-value limits of these processes, study their smoothness properties, and consider some interesting special cases. We further consider mixtures between such Cauchy processes and Gaussian processes, in order to have a separate control over the bulk and the tail dependence behaviors. We discuss inference methods for this class of processes, and show with a simulation study that our proposed inference approach yields accurate estimates. Moreover, the proposed class of models allows for a wide range of flexible dependence structures, and we demonstrate our new methodology by application to a temperature dataset.

Regime based Precipitation Modeling

Carolina Euan

Lancaster University
caroe@cimat.mx

Precipitation is one of the most important factors for different human activities. For instance, the presence or absence of rain can be crucial for agricultural and human settlements. Though, the intensity and duration of precipitation events can cause problems such as losing the harvest or flooding in central cities. Motivated by the versatility of regime based models, we proposed a hierarchical regime based spatiotemporal model for precipitation data. We consider the regime variable as a function of past values of the process in a neighborhood area. The spatial and temporal dependence is different among regimes. These regimes can be interpreted as the different rain systems observed in the data. We fit our model under the bayesian paradigm using INLA.

Session 60: Recent advances in statistical methods for modern degradation data

Planning Accelerated Degradation Tests with Two Stress Variables

♦*Guanqi Fang*[1]*, Rong Pan*[2] *and John Stufken*[3]
[1]Zhejiang Gongshang University
[2]Arizona State University
[3]University of North Carolina-Greensboro
gfang5@asu.edu

Conducting accelerated degradation tests (ADTs) is an effective way to assess reliability of highly reliable products. In the existing literature, most works deal with planning ADT with a single stress variable; however, the situation of more than one stress variable is commonly seen in engineering practice. To fill the gap, in this article, we provide an analytical approach to address the design issue when two stress variables are present. By using a linear mixed-effects model to describe the accelerated degradation process, we demonstrate that the design problem can be solved by, first, finding the optimal setting of test conditions and allocation of test units for a "single-variable" case, and then the initial solution is transformed to the test plan for the case of two stress variables. The transformation is done by maintaining the same value of the asymptotic variance of the estimated p-th quantile lifetime, along with the consideration of reducing the asymptotic variance of model parameters estimation. We also discuss how to find compromise plans that satisfy practical demands. Finally, the proposed framework is illustrated using a real-world example.

Accelerated degradation tests with inspection effects

Xiujie Zhao
Tianjin University
xiujiezhao@tju.edu.cn

This study proposes a framework to analyze accelerated degradation testing (ADT) data in the presence of inspection effects. Motivated by a real dataset from the electric industry, we study two types of effects induced by inspections. After each inspection, the system degradation level instantaneously reduces by a random value. Meanwhile, the degrading rate is elevated afterwards. Considering the absence of observations due to practical reasons, we employ the expectation-maximization (EM) algorithm to analytically estimate the unknown parameters in a stepwise Wiener degradation process with covariates. Moreover, to maintain the level of generality for the adaption of the method in various scenarios, a confidence density approach is utilized to hierarchically estimate the parameters in the acceleration link function. The proposed methods can provide efficient parameter estimation under complex link functions with multiple stress factors. Further, confidence intervals are derived based on the large-sample approximation. Simulation studies and a case study from Schneider Electric are used to illustrate the proposed methods. The results show that the proposed model yields a remarkably better fit to the Schneider data in comparison to the conventional Wiener ADT model.

A generalized Wiener process with dependent degradation rate and volatility and time-varying mean-to-variance ratio

Shirong Zhou[1]*, Yincai Tang*[1] *and* ♦*Ancha Xu*[2]
[1]East China Normal University
[2]Zhejiang Gongshang University
xuancha2011@aliyun.com

In degradation analysis, there exist two natural features for the degradation data. The first is the dependence of degradation rate and degradation volatility, and the other is the time-varying mean-to-variance ratio. Ignoring them may lead to a significant bias in assessing lifetime information of materials. This paper proposes a generalized Wiener degradation model, which puts these two characteristics and unit-to-unit variation into consideration simultaneously. The proposed model includes many existing Wiener degradation models as special cases. Then, a generalized closed-form approximated residual lifetime distribution is given for the proposed model. Statistical inference for the model parameters is conducted based on the expectation maximizing (EM) method, and two simple auxiliary frameworks are developed for the determination of initial values and the forms of the time-scaled functions. The developed methodologies are then illustrated and verified in a simulation study and two real data analysis.

Pairwise model discrimination with applications in lifetime distributions and degradation processes

♦*Piao Chen*[1]*, Zhisheng Ye*[2] *and Xun Xiao*[3]
[1]TU Delft
[2]National University of Singapore
[3]University of Otago
p.chen-6@tudelft.nl

Reliability data obtained from life tests and degradation tests have been extensively used for purposes such as estimating product reliability and predicting warranty costs. When there is more than one candidate model, an important task is to discriminate between the models. In the literature, the model discrimination was often treated as a hypothesis test and a pairwise model discrimination procedure was carried out. Because the null distribution of the test statistic is unavailable in most cases, the large sample approximation and the bootstrap were frequently used to find the acceptance region of the test. Although these two methods are asymptotically accurate, their performance in terms of size and power is not satisfactory in small sample size. To enhance the small-sample performance, we propose a new method to approximate the null distribution, which builds on the idea of generalized pivots. Conventionally, the generalized pivots were often used for interval estimation of a certain parameter or function of parameters in presence of nuisance parameters. In this study, we further extend the idea of generalized pivots to find the acceptance region of the model discrimination test. Through extensive simulations, we show that the proposed method performs better than the existing methods in discriminating between two lifetime distributions or two degradation models over a wide range of sample sizes. Two real examples are used to illustrate the proposed methods.

Session 61: Recent Developments in Meta-Analysis

A Bayesian model for combining standardized mean differences and odds ratios in the same meta-analysis

Yaqi Jing[1]*, Mohammad Hassan Murad*[2] *and* ♦*Lifeng Lin*[1]
[1]Florida State University
[2]Mayo Clinic
llin4@fsu.edu

In meta-analysis practice, researchers frequently face studies that report the same outcome differently, such as a continuous variable (e.g., scores for rating depression) or a binary variable (e.g., counts of patients with depression dichotomized by certain latent, unreported depression scores). To combine these two types of studies in the same analysis, a simple conversion method has been widely used to handle standardized mean differences (SMDs) and odds ratios

(ORs). This conventional method uses a linear function connecting the SMD and log OR, and it assumes logistic distributions for (latent) continuous measures. However, the normality assumption is more commonly used for continuous measures, and the conventional method may be inaccurate when effect sizes are large or cutoff values for dichotomizing binary events are extreme (leading to rare events). In this talk, we present a Bayesian hierarchical model to synthesize SMDs and ORs without using the conventional conversion method. This model assumes exact likelihoods for continuous and binary outcome measures, which account for full uncertainties in the synthesized results. We performed simulation studies to compare the performance of the conventional and Bayesian methods in various settings. The Bayesian method generally produced less biased results with smaller mean squared errors and higher coverage probabilities than the conventional method in most cases. We also used two case studies to illustrate the proposed Bayesian method in real-world settings.

Bivariate hierarchical Bayesian model for combining summary measures and their uncertainties from multiple sources

Yujing Yao[1]*, R. Todd Ogden*[1]*, Chubing Zeng*[2] *and* ♦*Qixuan Chen*[1]
[1]Columbia University
[2]University of Southern California
qc2138@cumc.columbia.edu

It is often of interest to combine available estimates of a similar quantity from multiple data sources. When the corresponding variances of each estimate are also available, a model should consider the uncertainty of the estimates themselves as well as the uncertainty in the estimation of variances. In addition, if there exists a strong association between estimates and their variances, the correlation between these quantities should also be considered. In this paper, we propose a bivariate hierarchical Bayesian model that jointly models the estimates and their estimated variances assuming a correlation between these two measures. We conduct simulations to explore the performance of the proposed bivariate Bayesian model and compare it to other commonly used methods under different correlation scenarios. The proposed bivariate Bayesian model has a wide range of applications. We illustrate its application in three very different areas: PET brain imaging studies, meta-analysis, and small area estimation.

Quantifying Replicability of Multiple Studies in a Meta-Analysis

♦*Mengli Xiao*[1]*, Haitao Chu*[1]*, James Hodges*[1] *and Lifeng Lin*[2]
[1]University of Minnesota
[2]Florida State University
xiaox345@umn.edu

For valid scientific discoveries, it is fundamental to evaluate whether research findings are replicable across different settings. While large-scale replication projects across broad research topics are not feasible, systematic reviews and meta-analyses (SRMAs) offer viable alternatives to assess replicability. Due to subjective inclusion and exclusion of studies, SRMAs may contain non-replicable studies. However, no rigorous statistical methods exist to characterize non-replicability in SRMAs. Non-replicability is often misconceived as high heterogeneity. This article introduces a new measure, the externally standardized residuals from a leave-study-out procedure, to quantify replicability. It not only measures the impact of non-replicability on the conclusion of an SRMA but also differentiates non-replicability from heterogeneity. A new test statistic for replicability is derived. We explore its asymptotic properties and use extensive simulations and three real-data studies to illustrate this measure's performance. We conclude that replicability should be routinely assessed for all SRMAs, and recommend sensitivity analyses once non-replicable studies are identified in an SRMA.

Bayesian inference for asymptomatic Covid-19 infection rates

Dexter Cahoy[1] *and* ♦*Joseph Sedransk*[2]
[1]University of Houston
[2]University of Maryland
jxs123@case.edu

To strengthen inferences, meta analyses are commonly used to summarize information from a set of independent studies. In some cases, though, the data may not satisfy the assumptions underlying the meta analysis. Using two Bayesian methods that have a more general structure than the common meta analytic ones, we can show the extent and nature of the pooling that is justified statistically. Here, we investigate by re-analyzing data from several reviews whose objective is to make inference about the Covid-19 asymptomatic infection rate. Our results show that it is unlikely that all of the true effect sizes come from a single source. Thus, researchers should be cautious about pooling the data from all of the studies.

Session 62: Modern streaming Data Analysis: Sequential Tests

On Sequential Test with Optimal Observation Strategy for Detection of Change in Sensor Networks

♦*Dong Han*[1]*, Fugee Tsung*[2] *and Lei Qiao*[1]
[1]Shanghai Jiao Tong University
[2]Hong Kong University of Science and Technology
donghan@sjtu.edu.cn

This paper considers a decentralized sequential detection problem of a sensor network in which each sensor receives a sequence of observations and sends a sequence of observed information to a fusion center where a final decision about the change in distribution of the network is made quickly by a sequential test, subject to constraints on the false alarm rate of the center, the average number of observations of each sensor. A new measure of detection delay and an optimal sequential test of the center with optimal observation strategy of each sensor to detect the change in distribution of the network, are proposed for finite Markov observation sequence of each sensor. Moreover, the numerical simulations is provided to illustrate the theoretical results.

Adversarially Robust Sequential Hypothesis Testing

Shuchen Cao[1]*,* ♦*Ruizhi Zhang*[1] *and Shaofeng Zou*[2]
[1]University of Nebraska-Lincoln
[2]University 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 few 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. Specifically, under the null hypothesis, the adversary

picks a distribution $p \in \mathcal{P}$ with cost $c_0(p)$; and under the alternative hypothesis, the adversary picks a distribution $q \in \mathcal{Q}$ with cost $c_1(q)$. 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.

A multistage test for high-dimensional signal recovery

Yiming Xing and ♦*Georgios Fellouris*
University of Illinois, Urbana-Champaign
fellouri@illinois.edu

A multistage test, in which sampling can be terminated at 4 fixed times, is proposed for the fundamental hypothesis testing problem. The average sample size required by this testing procedure under both hypotheses is the same as the optimal to a first-order asymptotic approximation as the target type-I and type-II error probabilities go to zero. At the same time, the proposed test is substantially more robust than the Sequential Probability Ratio Test when the true hypothesis is in the indifference zone. This testing approach is then applied to the multiple testing problem where a distinct test must be applied to each hypothesis. The asymptotic optimality of the resulting procedure is established under certain sparsity rates as the number of hypotheses goes to infinity and the target familywise error rates remain fixed. The exact performance of this test is compared in concrete cases against that of existing multistage multiple testing procedures in the literature.

Asymptotically optimal sequential FDR and pFDR control

Jay Bartroff
University of Southern California
bartroff@usc.edu

I will discuss asymptotically optimal multiple testing procedures for sequential data in the context of prior information on the number of false null hypotheses, for controlling FDR/FNR or pFDR/pFNR. These procedures are closely related to those proposed and shown by Song & Fellouris (2017, Electron. J. Statist.) to be asymptotically optimal for controlling type 1 and 2 familywise error rates (FWEs). Further, we show that by appropriately adjusting the critical values of the Song-Fellouris procedures, they can be made asymptotically optimal for controlling any multiple testing error metric that is bounded between multiples of FWE in a certain sense. In addition to FDR/FNR and pFDR/pFNR this includes other metrics like the per-comparison and per-family error rates, and the false positive rate. Our asymptotic analysis includes regimes in which the number of null hypotheses approaches infinity.

Session 63: Modern Statistical Methods for Complex Data Objects Analysis

Unified Principal Component Analysis for Sparse and Dense Functional Data under Spatial Dependency

Yehua Li
University of California at Riverside
yehuali@ucr.edu

We consider spatially dependent functional data collected under a geostatistics setting, where spatial locations are irregular and random. The functional response is the sum of a spatially dependent functional effect and a spatially independent functional nugget effect. Observations on each function are made on discrete time points and contaminated with measurement errors. Under the assumption of spatial stationarity and isotropy, we propose a tensor product spline estimator for the spatio-temporal covariance function. When a coregionalization covariance structure is further assumed, we propose a new functional principal component analysis method that borrows information from neighboring functions. The proposed method also generates nonparametric estimators for the spatial covariance functions, which can be used for functional kriging. Under a unified framework for sparse and dense functional data, infill and increasing domain asymptotic paradigms, we develop the asymptotic convergence rates for the proposed estimators. Advantages of the proposed approach are demonstrated through simulation studies and two real data applications representing sparse and dense functional data, respectively.

Spline smoothing of 3D geometric data

♦*Xinyi Li*[1]*, Shan Yu*[2]*, Yueying Wang*[3]*, Guannan Wang*[4] *and Lily Wang*[5]
[1]Clemson Univeristy
[2]University of Virginia
[3]Columbia University
[4]College of William and Mary
[5]George Mason University
lixinyi@clemson.edu

Over the past two decades, we have seen an increased demand for 3D visualization and simulation software in medicine, architectural design, engineering, and many other areas, which have boosted the investigation of geometric data analysis and raised the demand for further advancement in statistical analytic approaches. We propose a class of spline smoothers appropriate for approximating geometric data over 3D complex domains, which can be represented in terms of a linear combination of spline basis functions with some smoothness constraints. We start with introducing the tetrahedral partitions, Barycentric coordinates, Bernstein basis polynomials, and trivariate spline on tetrahedra. Then, we propose a penalized spline smoothing method for identifying the underlying signal in a complex 3D domain from potential noisy observations. Simulation studies are conducted to compare the proposed method with traditional smoothing methods on 3D complex domains.

Sequential Change-point Detection for High-Dimensional and non-Euclidean Data

♦*Lynna Chu*[1] *and Hao Chen*[2]
[1]Iowa State University
[2]University of California, Davis
lchu@iastate.edu

In many modern applications, high-dimensional/non-Euclidean data sequences are collected to study complex phenomena over time and it is often of scientific significance to detect anomaly events as data is continually being collected. We study a nonparametric framework that utilizes nearest neighbor information among the observations and can be applied to various data types to detect changes in an online setting. We consider new test statistics under this framework that can detect anomaly events more effectively than the existing test with the false discovery rate controlled at the same level. Analytical formulas to determine the threshold of claiming a change are also provided, making the approach easily applicable for real data applications.

Broadcasted Nonparametric Tensor Regression

Ya Zhou[1]*,* ♦*Raymond K. W. Wong*[2] *and Kejun He*[3]
[1]Renmin University of China and Texas A&M University
[2]Texas A&M University

[3]Renmin University of China
raywong@tamu.edu
We propose a novel broadcasting idea to model the nonlinearity in tensor regression non-parametrically. Unlike existing nonparametric tensor regression models, the resulting model strikes a good balance between flexibility and interpretability. In this talk, I will discuss the proposed estimation method, and the corresponding theoretical investigation. Empirical results will also be presented.

Session 64: On inference for time series models

Location-adaptive change-point testing for time series
Linlin Dai and ♦*Rui She*
Southwestern University of Finance and Economics
rshe@swufe.edu.cn
We propose a location-adaptive self-normalization (SN) based test for change points in time series. The SN technique has been extensively used in change point detection for its capability to avoid direct estimation of nuisance parameters. Compared with the traditional Kolmogorov-Smirnov test, the SN-based tests are known to have a better size and deliver a monotonic power. However, we find that the power of the SN-based test is susceptible to the location of the break and may suffer from a severe power loss, especially when the change occurs at the early or late stage of the sequence. This phenomenon is essentially caused by the unbalance of the data used before and after the change point when one is building a test statistic based on the cumulative sum (CUSUM) process. Hence, we consider leaving out the samples far away from the potential locations of change points and propose an optimal data selection scheme. Based on this scheme, a new SN-based test statistic adaptive to the locations of breaks is established. The new test can significantly improve the power of the existing SN-based tests while maintaining a satisfactory size. It is a unified treatment that can be readily extended to tests for general quantities of interest, such as the median and the model parameters. To our knowledge, we are the first to bring a location-adaptive idea into the change point detection. The derived optimal subsample selection strategy is not specific to the SN-based tests but is applicable to any method that relies on the CUSUM process, and it may provide new insights in the area for future research.

Statistical inferences for threshold GARCH model based on the intraday high frequency data
Xingfa Zhang
Guangzhou University
xingfazhang@gzhu.edu.cn
ased on the scale model framework, intraday high frequency data is introduced to improve the estimation and test of the threshold GARCH model. New model estimator and test statistic for adequacy using the high frequency data information are proposed. Simulations show that the new estimator can have significantly smaller standard deviation compared to the traditional one, and the new test statistic can have better power. A practical study of CSI300 index is given to show the potential applications of the proposed approach.

Bootstrapping on robust goodness-of-fit test for GARCH models
♦*Xuqin Wang and Muyi Li*
Xiamen University
xqwang_0408@163.com
We consider a random weighting bootstrap approach to approximate the empirical distribution function-based portmanteau test for generalized autoregressive conditional heteroskedastic (GARCH) models estimated by the least absolute deviation (LAD) method. The test is applicable for very heavy-tailed innovations with only finite fractional moments. Besides, the random weighting bootstrap method is easy to implement and insensitive to the choice of the perturbing random weighting, thus no user-chosen parameters are needed. Simulations will be conducted to assess the finite sample performance of the proposed test, and the analysis of real data will be given to illustrate the usefulness of the test.

Smooth transition moving average models: estimation, testing, and computation
♦*Xinyu Zhang and Dong Li*
Tsinghua University
zhang-xy17@mails.tsinghua.edu.cn
This paper introduces a new class of nonlinear moving average model, called the smooth transition moving average (STMA) model, and studies its least squares estimation (LSE). It is shown that, under some mild conditions, the LSE is strongly consistent and asymptotically normal. A powerful score-based goodness-of-¡U+FB01¿t test for the MA and STMA model is presented. A different parameterization from the classical one is applied to numerically improve the identification and estimation of this model. Simulation studies are conducted to assess the performance of the LSE and the score-based test in ¡U+FB01¿nite samples. The results are illustrated with an application to the weekly exchange rate of the USA Dollar to the British Pound from 1971 to 2020.

Session 65: recent advances in nonparametric and robust estimation and inference for complex data

New Regression Model: Modal Regression
♦*Weixin Yao*[1]*, Longhai Li*[2] *and Sijia Xiang*[3]
[1]University of California, Riverside
[2]University of Saskatchewan
[3]Zhejiang University of Finance & Economics
weixin.yao@ucr.edu
Built on the ideas of mean and quantile, mean regression and quantile regression are extensively investigated and popularly used to model the relationship between a dependent variable Y and covariates x. However, the research about the regression model built on the mode is rather limited. In this talk, we propose a new regression tool, named modal regression, that aims to find the most probable conditional value (mode) of a dependent variable Y given covariates x rather than the mean that is used by the traditional mean regression. The modal regression can reveal new interesting data structure that is possibly missed by the conditional mean or quantiles. In addition, modal regression is resistant to outliers and heavy-tailed data, and can provide shorter prediction intervals when the data are skewed. Furthermore, unlike traditional mean regression, the modal regression can be directly applied to the truncated data. Modal regression could be a potentially very useful regression tool that can complement the traditional mean and quantile regressions.

Big spatial data learning: a parallel solution
Shan Yu[1]*, Guannan Wang*[2] *and* ♦*Lily Wang*[3]
[1]University of Virginia
[2]College of William and Mary
[3]George Mason University
lwang41@gmu.edu
We are living in the era of ”Big Data.” A significant portion of big data is, in fact, big spatial data captured through remote sensors, ad-

vanced technologies or large-scale simulations. Explosive growth in the spatial and/or spatiotemporal data emphasizes the need for developing new and computationally efficient methods tailored for analyzing such large-scale data. Parallel statistical computing has proved to be a handy tool when dealing with big data. In general, it uses multiple processing elements simultaneously to solve a problem. Some statistical problems are naturally parallel and can be easily decomposed into independent tasks, such as the Monte Carlo integration, Multiple chains of MCMC, and Bootstrap for confidence intervals. However, most of the statistical program segments are not independent, and thus they are difficult to be executed in parallel, such as the conventional spline regression. In this work, we develop a novel parallel smoothing technique based on multivariate spline over triangulation, which can be used under different hardware parallelism levels. Moreover, we show that the parallelized estimators reach the same convergence rate as the global estimators. In addition, conflated with concurrent computing, the proposed method can be easily extended to the distributed system.

Spatial Homogeneity Regression: A Bayesian Nonparametric Recourse

Guanyu Hu
University of Missouri
guanyu.hu@missouri.edu

Spatial regression models are ubiquitous in many different areas such as environmental science, geoscience, and public health. Exploring relationships between response variables and covariates with complex spatial patterns is a very important work. In this talk, we will discuss a novel spatially clustered coefficients regression model based on nonparametric Bayesian methods. The theoretical properties of our proposed method are established. An efficient Markov chain Monte Carlo algorithm is designed. Extensive simulation studies are carried out to examine empirical performance of the proposed method. Additionally, we demonstrate the excellent empirical performance of our method via various applications.

BEAUTY powered BEAST

♦*Kai Zhang*[1]*, Zhigen Zhao*[2] *and Wen Zhou*[3]
[1]UNC Chapel Hill
[2]Temple University
[3]Colorado State University
zhangk@email.unc.edu

We study inference about the uniform distribution with the proposed binary expansion approximation of uniformity (BEAUTY) approach. Through an extension of the celebrated Euler's formula, we approximate the characteristic function of any copula distribution with a linear combination of means of binary interactions from marginal binary expansions. This novel characterization enables a unification of many important existing tests through an approximation from some quadratic form of symmetry statistics, where the deterministic weight matrix characterizes the power properties of each test. To achieve a uniformly high power, we study test statistics with data-adaptive weights through an oracle approach, 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 leads all existing tests we considered in empirical power against all complex forms of alternatives. This oracle therefore sheds light on the potential of substantial improvements in power and on the form of optimal weights under each alternative. By approximating this oracle with data-adaptive weights, we develop the BEAST that improves the empirical power of many existing tests against a wide spectrum of common alternatives while providing clear interpretation of the form of non-uniformity upon rejection. We illustrate the BEAST with a study of the relationship between the location and brightness of stars.

Session 66: Novel Methods for Statistical Analysis of Complex Data

Semiparametric Regression Models for Between- and Within-subject Attributes: Asymptotic Efficiency and Applications to High-Dimensional Data

Xin Tu
Division of Biostat and Bioinfo, UCSD
X2tu@ucsd.edu

Breakthroughs in high-throughput sequencing technologies have facilitated the understanding of inherent disease mechanisms, while high-dimensional sequence data also provoke substantial challenges in statistical analyses and interpretations. As an interesting way to look at such high dimensional sequencing data, between-subject attributes comparing two subjects' high-dimensional data such as genome sequences using dissimilarity/similarity metrics are now widely used as dimension-reduction summary outcomes. To derive meaningful relationships and interpretations between such outcomes and clinical phenotypes, one needs to not only address statistical challenges arising from mixed within- and between-subject attributes but also handle correlations among the between-subject attributes. In this talk, we discuss new semiparametric regression models for relationships of an outcome (either between- or within-subject) with explanatory variables (between- and within-subject). In the burgeoning fields of high-throughput data, the proposed semiparametric approach fills a critical methodological gap by deriving robust regression relationships between mixed low- and high-dimensional outcomes. Unlike existing regularized regression models, the proposed approach can not only model high-dimensional outcomes as a response (or dependent variable), but also allow such outcomes to have any dimension. Like semiparametric regression models for within-subject attributes, the proposed semiparametric models with inference based on a set of U-statistics based generalized estimating equations are also semiparametric efficient. We illustrate the approach with both simulated and real study data in microbiome research, where dimensions are in the tens of thousands.

fMRI data classification based on the nonparametric bayesian graph model

Chong Wan and ♦*Guoxin Zuo*
Central China Normal University
zuogx@mail.ccnu.edu.cn

Modeling functional brain unweighted network by a non-parametric generalized stochastic block model which called Infinite Relation Model(IRM) is practical. But it is not adequate for functional brain weighted network. In this paper, we construct weighted functional brain network, and propose a generalized IRM is to deal with weighted network. inferring this model by MCMC, and checking the model's fitness or predictive performance. In simulated data, generalized IRM can get almost all hidden node cluster , better than IRM, and also shows better performance in fitness and prediction for functional brain weighted network data.

Diagnosis of thyroid nodules for ultrasonographic characteristics indicative of malignancy using random forest

Dan Chen[1]*,* ♦*Jun Hu*[2]*, Mei Zhu*[3]*, Niansheng Tang*[1]*, Yang Yang*[3]

and Yuran Feng[3]
[1]Yunnan University
[2]Yunnan Agricultural University
[3]The First Affiliated Hospital of Kunming Medical University
1995006@ynau.edu.cn

Various combinations of ultrasonographic (US) characteristics are increasingly utilized to classify thyroid nodules. But they lack theories, heavily depend on radiologists' experience, and cannot correctly classify thyroid nodules. Hence, our main purpose is to develop an efficient scoring system for facilitating ultrasonic clinicians to correctly identify thyroid malignancy. We calculate the probability for each of thyroid nodules via random forest (RF) and extreme learning machine (ELM), and develop a scoring system to classify thyroid nodules. For comparison, we also consider eight state-of-the-art methods such as support vector machine (SVM), neural network (NET), etc. The area under the receiver operating characteristic curve (AUC) is employed to measure the accuracy of various classifiers. Using the developed scoring system, thyroid nodules are classified into the following four categories: benign, low suspicion, intermediate suspicion, and high suspicion, whose rates of malignancy correctly identified for RF (ELM) method on the testing dataset are 0.0% (4.3%), 14.3% (50.0%), 58.1% (59.1%) and 96.1% (97.7%), respectively. The random forest performs better than other methods in identifying malignancy, especially for abnormal nodules, in terms of risk scores. The developed scoring system can well predict the risk of malignancy and guide medical doctors to make management decisions for reducing the number of unnecessary biopsies for benign nodules.

A class of weighted estimating equations for additive hazard models with covariates missing at random

PENG YE
School of Statistics, University of International Business and Economics, Beijing 100029, China
yepeng@uibe.edu.cn

Missing covariate data arise frequently in biomedical studies. In this article, we propose a class of weighted estimating equations for the additive hazard regression model when some of the covariates are missing at random. Time-specific and subject-specific weights are incorporated into the formulation of weighted estimating equations. Unified results are established for estimating selection probabilities that cover both parametric and non-parametric modeling schemes. The resulting estimators have closed forms and are shown to be consistent and asymptotically normal. Simulation studies indicate that the proposed estimators perform well for practical settings. An application to a mouse leukemia study is illustrated.

Session 67: Advances in Models and Methods for Complex Spatial Processes

The frequency and severity of crop damage by wildlife in rural Beijing, China

♦*Liang Fang*[1]*, Yiping Hong*[2]*, Zaiying Zhou*[3] *and Wenhui Chen*[1]
[1]Beijing Forestry University
[2]King Abdullah University of Science and Technology
[3]Tsinghua University
tsinghuafl@163.com

With a continuous expansion of the number and activity of wild animals induced by ecological conservation and restoration efforts, the human-wildlife conflict is becoming more prominent across China. There have been frequent and severe incidents of crop damage caused by wildlife. In this talk, we investigate the crop losses caused by wildlife in the rural districts of Beijing, using a unique dataset of 31,573 observations from 2009 to 2017. Through statistical tests on the individual coefficients and the overall fitness, we find that a negative binomial generalized regression model describes the pattern of crop loss events more accurately, compared to an alternative Poisson model. The frequency of crop loss events is positively related to a village's distance from the river system but negatively associated with the distance from woodland, population density, and protective measures taken. The predicted frequencies of crop damage events are then used to correlate with the amounts of losses at the village level. Based on these results, we propose solutions for effective reduction of future crop losses and practical assessment of the likely damage compensation or insurance premium.

Copula-based Multiple Indicator Kriging for non-Gaussian Random Fields

Gaurav Agarwal
Statistics at Lancaster University, UK
g.agarwal@lancaster.ac.uk

In spatial statistics, the kriging predictor is the best linear predictor at unsampled locations, but not the optimal predictor for non-Gaussian processes. In this research, we introduce a copula-based multiple indicator kriging model for the analysis of non-Gaussian spatial data by thresholding the spatial observations at a given set of quantile values. The proposed copula model allows for flexible marginal distributions while modeling the spatial dependence via copulas. We show that the covariances required by kriging have a direct link to the chosen copula function. We then develop a semiparametric estimation procedure. The proposed method provides the entire predictive distribution function at a new location, and thus allows for both point and interval predictions. The proposed method demonstrates better predictive performance than the commonly used variogram approach and Gaussian kriging in the simulation studies. We illustrate our methods on precipitation data in Spain during November 2019, and heavy metal dataset in topsoil along the river Meuse, and obtain probability exceedance maps.

Efficient Calibration of Numerical Model Output Using Hierarchical Dynamic Models

♦*Yewen Chen*[1]*, Xiaohui Chang*[2] *and Hui Huang*[1]
[1]Sun Yat-sen university
[2]Oregon State University
chenyw68@mail2.sysu.edu.cn

Numerical air quality models, such as the Community Multiscale Air Quality (CMAQ) system, play a critical role in characterizing pollution levels at fine spatial and temporal scales. Nevertheless, numerical model outputs may systematically overestimate or underestimate pollutants concentrations due to various reasons. In this work, we propose an hierarchal dynamic model that can be implemented to calibrate grid-level CMAQ outputs using point-level observations from sparsely located monitoring stations. Under a Bayesian framework, our model presents a flexible quantification of uncertainties by considering deep hierarchies for key parameters, which can also be used to describe the dynamic nature of data structural changes. In addition, we adopt several newly-emerged techniques, including triangulations of research domain, tapering-based Gaussian kernel, sparse Gaussian graphical model, variational Bayes and ensemble Kalman smoother, that significantly speed up the whole calibration procedure. Our approach is illustrated using daily PM2.5 datasets of China's Beijing-Tianjin-Hebei region, which contains 68 monitoring stations and is covered by 2499 CMAQ 9-km grids. In contrast to existing methods, our model

gives more accurate calibration results in most of the grids with higher computation efficiencies. This allows us to provide an effective calibration tool for large-scale numerical model outputs and generate better high-resolution maps of pollutants.

A Nonstationary Spatio-Temporal Autologistic Regression Model

♦*Yiping Hong, Marc G. Genton and Ying Sun*
King Abdullah University of Science and Technology
yiping.hong@kaust.edu.sa

In many research fields such as ecology and epidemiology, the spatio-temporal datasets are binary, describing the existence of particular species or events. The spatial dependence of the binary observations often exhibits non-stationarity. However, most existing non-stationary models are based on Gaussian random fields, which are inappropriate for binary data. We propose a non-stationary spatio-temporal autologistic regression model, which allows the spatial covariances to vary in space. We investigate the spatial and temporal correlation of autologistic models with different coding and centering settings, suggesting the non-centered $\{-1, 1\}$ coding model. We then develop the maximum pseudolikelihood method for parameter estimation, prediction, and model selection. The simulation studies show the superior performance of the proposed methods compared to commonly used models. We apply our model to analyze the binary data of the particular matter ($PM_{2.5}$) in China thresholding at the health standard value. The results describe the non-stationarity in spatial dependence across China as well as predict the risk probability of the $PM_{2.5}$.

Session 68: Recent development in complex dependent data analysis

Spiked Eigenvalues of High-dimensional Sample Auto-covariance Matrices: CLT and Applications

Daning Bi[1]*, Xiao Han*[2]*, Adam Nie*[1] *and* ♦*Yanrong Yang*[1]
[1]The Australian National University
[2]University of Science and Technology of China
yanrong.yang@anu.edu.au

Based on factor modeling on high dimensional time series, this paper contributes to asymptotic properties of spiked eigenvalues for high dimensional sample auto-covariance matrices. The central limit theorem (CLT) is established under a general scenario in the sense of the following two aspects: 1. moderate high dimensional setting where the dimension p and the sample size T are comparable; 2. allowing time lag to be fixed or tend to infinity. Based on this CLT, a novel auto-covariance equivalence test is proposed for two independent high-dimensional time series. This test compares spiked eigenvalues of two high dimensional time series under study and facilitates further statistical applications such as clustering multiple-population high-dimensional time series. Various simulations demonstrate excellent performance of the proposed test. An empirical analysis on hierarchical clustering for multiple-country mortality data is conducted.

Quantile index regression

♦*Yingying Zhang*[1]*, Jingyu Zhao*[2]*, Guodong Li*[2] *and Chil-Ling Tsai*[3]
[1]East China Normal University
[2]University of Hong Kong
[3]University of California at Davis
yyzhang@fem.ecnu.edu.cn

It is usually difficult for quantile regression to make inference at levels with sparse observations, such as high and low quantiles, and the situation becomes more serious for high-dimensional data. This paper solves the problem by alterna- tively conducting the estimating procedure at levels with rich observations and then extrapolating the

fitted structures to high or low levels. It leads to a novel partially parametric quantile regression model, called the quantile index regression. Asymp- totic properties are derived for the case with

fixed dimensions, and non-asymptotic error bounds are established for high-dimensional data. Simulation experiments and a real data example demonstrate the usefulness of the proposed model over various existing methods.

SARMA: A Novel Computationally Scalable High-Dimensional Vector Autoregressive Moving Average Model

♦*Feiqing Huang*[1]*, Yao Zheng*[2]*, Di Wang*[3] *and Guodong Li*[1]
[1]University of Hong Kong
[2]University of Connecticut
[3]University of Chicago
u3514252@connect.hku.hk

Classical VARMA models are very popular in modeling the general linear process due to their model parsimony and good forecasting performance, yet the complicated idenfication issues and heavy computational burdens hinder their practicality in the high dimensional regime. In this paper, we introduce a scalable autoregressive moving average (SARMA) model that inherits the interpretability and rich dynamic structure of the VARMA models, while avoiding the identification problem. Most notably, our proposed model can be easily extended to include not only all the VARMA processes, but also other forms of general linear processes. We establish the non-asymptotic error bounds for the SARMA model, and further propose the first computationally scalable algorithm for VARMA estimation in high dimension. Simulation and real data experiments are given to demonstrate the advantages of the proposed approach over various existing methods.

Robust estimation of high-dimensional vector autoregressive models

♦*Di Wang and Ruey S. Tsay*
University of Chicago
di.wang@chicagobooth.edu

High-dimensional time series data appear in many scientific areas under the current data-rich environment. Analysis of such data poses new challenges to data analysts because of not only the complicated dynamic dependence between the series, but also the existence of aberrant observations, such as missing values, contaminated observations, and heavy-tailed distributions. For high-dimensional vector autoregressive (VAR) models, we introduce a unified estimation procedure that is robust to model misspecification, heavy-tailed noise contamination, and conditional heteroscedasticity. The proposed methodology enjoys both statistical optimality and computational efficiency, and can handle many popular high-dimensional models, such as sparse, reduced-rank, banded, and network-structured VAR models. With proper regularization and data truncation, the estimation convergence rates are shown to be nearly optimal under a bounded fourth moment condition. Consistency of the proposed estimators is also established under a relaxed bounded (2 + 2e)-th moment condition, for some e ¡U+2208¿ (0, 1), with slower convergence rates associated with e. The efficacy of the proposed estimation methods is demonstrated by simulation and a real example.

Panel discussion: Statistics and Data Science Partnerships and Collaborations across Sectors

Statistics and Data Science Partnerships and Collaborations across Sectors

Kelly Zou[1]*, John Kolassa*[2]*, Jim Li*[1]*, Fanni Natanegara*[3]*, Kimberly Sellers*[4] *and* ♦*Aniketh Talwai*[5]
[1]Viatris
[2]Rutgers, The State University of New Jersey
[3]Eli Lilly & Company
[4]Georgetown University
[5]Medidata - a Dassault Systems Company
atalwai@mdsol.com
Already submitted by Kelly Zou, Viatris

Statistics and Data Science Partnerships and Collaborations across Sectors

Kelly Zou[1]*,* ♦*Kimberly Sellers*[2]*, John Kolassa*[3]*, Jim Li*[1] *and Fanni Natanegara*[4]
[1]Viatris
[2]Georgetown U.
[3]Rutgers U.
[4]Eli Lilly and Company
kfs7@georgetown.edu
N/A

Statistics and Data Science Partnerships and Collaborations across Sectors

John Kolassa[1]*, Jim Li*[2]*, Fanni Natanegara*[3]*, Kimberly Sellers*[4] *and* ♦*Kelly Zou*[5]
[1]kolassa@stat.rutgers.edu
[2]Jim.Li@viatris.com
[3]natanegara$_{fanni@lilly.com}$
[4]Kimberly.Sellers@georgetown.edu
[5]Kelly.Zou@viatris.com
KelZouDS@gmail.com
Collaborations and partnerships can come in all shapes and forms. Martin Luther King, Jr.'s words may resonate: "We may have all come on different ships, but we're in the same boat now." Frequently, sharing of ideas between stakeholders from different organizations leads to exchange visits, support for graduate students, consulting jobs, grant support, and continuing education opportunities for statisticians or data scientists outside of academe. Although these activities statistical and data science problems from outside academia become use cases studies. The intellectual exchange that results is a key component of such partnerships. This panel includes experts from industry and consulting, which will have a wide appeal, given the increasing focus on inter-disciplinary research and the emergence of complex and high dimensional data. In particular, such challenges are common in health care research. In this invited 90-minute session, several panelists discuss the key elements to form and sustain successful collaborations and partnerships, along with challenges and barriers. The panel discussion can be valuable to statisticians and data scientists in diverse areas and sectors. (Keywords: Collaboration; Partnership; Career Sector; Data Science Statistics; Multidisciplinary Research)

Statistics and Data Science Partnerships and Collaborations across Sectors

Victoria Gamerman[1]*, John Kolassa*[2]*, Jim Li*[3]*,* ♦*Fanni Natanegara*[4]*, Kimberly Sellers*[5]*, Aniketh Talwai*[6]*, Aniketh Talwai*[6]*, Aniketh Talwai*[6]*, Aniketh Talwai*[6] *and Aniketh Talwai*[6]
[1]Boehringer Ingelheim
[2]Rutgers, The State University of New Jersey
[3]Viatris
[4]Eli Lilly and Company
[5]Georgetown University
[6]Medidata, a Dassault Systèmes company
natanegara_fanni@lilly.com
Collaborations and partnerships can come in all shapes and forms. Martin Luther King, Jr.'s words may resonate: "We may have all come on different ships, but we're in the same boat now." Frequently, sharing of ideas between stakeholders from different organizations leads to exchange visits, support for graduate students, consulting jobs, grant support, and continuing education opportunities for statisticians or data scientists outside of academe. Although these activities statistical and data science problems from outside academia become use cases studies. The intellectual exchange that results is a key component of such partnerships. This panel includes experts from industry and consulting, which will have a wide appeal, given the increasing focus on inter-disciplinary research and the emergence of complex and high dimensional data. In particular, such challenges are common in health care research. In this invited 90-minute session, several panelists discuss the key elements to form and sustain successful collaborations and partnerships, along with challenges and barriers. The panel discussion can be valuable to statisticians and data scientists in diverse areas and sectors. Keywords: Collaboration; Partnership; Career Sector; Data Science Statistics; Multidisciplinary Research

Session 69: Novel statistical models of -omic data analysis

DiSNEP: a Disease-Specific gene Network Enhancement to improve Prioritizing candidate disease genes

Peifeng Ruan[1] *and* ♦*Shuang Wang*[2]
[1]Yale University
[2]Columbia University
sw2206@cumc.columbia.edu
Biological network-based strategies are useful in prioritizing genes associated with diseases. Several comprehensive human gene networks such as STRING, GIANT and HumanNet were developed and used in network-assisted algorithms to identify disease-associated genes. None of them are disease-specific and may not accurately reflect gene interactions for a specific disease. Aiming to improve disease gene prioritization using networks, we propose a Disease-Specific Network Enhancement Prioritization (DiSNEP) framework. DiSNEP enhances a comprehensive gene network for a disease through a diffusion process on a gene-gene similarity matrix derived from a disease omics data. The enhanced disease-specific gene network thus better reflects true gene interactions for the disease and improves prioritizing disease-associated genes subsequently. In simulations, DiSNEP prioritizes more true signal genes than competing methods using a general gene network. Applications to prioritize cancer-associated gene expression and DNA methylation signal genes for five cancer types from The Cancer Genome Atlas project suggest that more prioritized candidate genes by DiSNEP are cancer-related according to the DisGeNET database consistently across all five cancer types considered.

EPIC: inferring relevant cell types for complex traits by integrating genome-wide association studies and single-cell RNA sequencing

Rujin Wang, Danyu Lin and ♦*Yuchao Jiang*
University of North Carolina at Chapel Hill
yuchaoj@email.unc.edu

More than a decade of genome-wide association studies (GWASs) have identified genetic risk variants that are significantly associated with complex traits. Emerging evidence suggests that the function of trait-associated variants likely acts in a tissue- or cell-type-specific fashion. Yet, it remains challenging to prioritize trait-relevant tissues or cell types to elucidate disease etiology. Here, we present EPIC (cEll tyPe enrIChment), a statistical framework that relates large-scale GWAS summary statistics to cell-type-specific gene expression measurements from single-cell RNA sequencing (scRNA-seq). We derive powerful gene-level test statistics for common and rare variants, separately and jointly, and adopt generalized least squares to prioritize trait-relevant cell types while accounting for the correlation structures both within and between genes. Using enrichment of loci associated with four lipid traits in the liver and enrichment of loci associated with three neurological disorders in the brain as ground truths, we show that EPIC outperforms existing methods. We apply our framework to multiple scRNA-seq datasets from different platforms and identify cell types underlying type 2 diabetes and schizophrenia. The enrichment is replicated using independent GWAS and scRNA-seq datasets and further validated using PubMed search and existing bulk case-control testing results.

Robust estimation of cell type fractions from bulk data via ensemble of various deconvolution approaches

♦*Jiebiao Wang, Manqi Cai and Wei Chen*
University of Pittsburgh
jbwang@pitt.edu

Omics analyses at tissue level are known to be confounded by cell type heterogeneity. To adjust for this, many statistical methods, which are called cell type deconvolution, have been proposed to infer cell-type fractions from bulk/tissue-level data. However, these methods produce vastly different results under various settings. Benchmarking also shows no universally best combination of different factors in deconvolution performance. To achieve robust estimation of cell type fractions, we propose using ensemble estimation to incorporate and unify the results from top existing deconvolution methods, reference datasets, marker gene selection procedures, data transformations, and normalizations. Using several large real datasets with ground truth of measured cell type fractions, we evaluate the proposed method's performance in different tissue types and demonstrate that our method yields more stable and accurate results than existing deconvolution methods. Additionally, ensemble deconvolution can identify cell types with differential fractions associated with clinical outcomes. In the end, we will discuss the downstream cell-type-specific analyses enabled by robust estimation of cell type fractions.

Bayesian Genome-wide TWAS method integrating both cis- and trans- eQTL with GWAS summary statistics

Justin Luningham[1]*, Junyu Chen*[2]*, Shizhen Tang*[2]*, Philip De Jager*[3]*, David Bennett*[4]*, Aron Buchman*[4] *and* ♦*Jingjing Yang*[2]
[1]Georgia State University
[2]Emory University
[3]Columbia University Irving Medical Center
[4]Rush University Medical Center
jingjing.yang@emory.edu

Transcriptome-wide association studies (TWAS) have been widely used to integrate gene expression and genetic data for studying complex traits. Due to the computational burden, existing TWAS methods do not assess distant trans- expression quantitative trait loci (eQTL) that are known to explain important expression variation for most genes. We propose a Bayesian Genome-wide TWAS (BGW-TWAS) method which leverages both cis- and trans- eQTL information for TWAS. Our BGW-TWAS method is based on Bayesian variable selection regression, which not only accounts for cis- and trans- eQTL of the target gene but also enables efficient computation by using summary statistics from standard eQTL analyses. Our simulation studies illustrated that BGW-TWAS achieved higher power compared to existing TWAS methods that do not assess trans-eQTL information. We further applied BWG-TWAS to individual-level GWAS data (N= 3.3K), which identified significant associations between the genetically regulated gene expression (GReX) of gene ZC3H12B and Alzheimer's dementia (AD) (p-value=5.42 e-13), neurofibrillary tangle density (p-value=1.89e-6), and global measure of AD pathology (p-value=9.59e-7). These associations for gene ZC3H12B were completely driven by trans-eQTL. Additionally, the GReX of gene KCTD12 was found to be significantly associated with ß-amyloid (p-value=3.44e-8) which was driven by both cis- and trans- eQTL. Four of the top driven trans-eQTL of ZC3H12B are located within gene APOC1, a known major risk gene of AD and blood lipids. Additionally, by applying BGW-TWAS with summary-level GWAS data of AD (N= 54K), we identified 13 significant genes including known GWAS risk genes HLA-DRB1 and APOC1, as well as ZC3H12B.

Session 70: Statistical methods for spatial genomics and transcriptomics

Statistical methods for spatial genomics and transcriptomics

♦*Mingyao Li, Jian Hu, Kyle Coleman and Amelia Schroeder*
University of Pennsylvania
mingyao@pennmedicine.upenn.edu

Recent developments in spatial transcriptomics technologies have enabled scientists to get an integrated understanding of cells in their morphological context. Applications of these technologies in diverse tissues and diseases have transformed our views of transcriptional complexity.

Learning fine-resolution Hi-C contact maps

Wenxiu Ma
University of California Riverside
wenxiu.ma@ucr.edu

High-throughput methods based on chromosome conformation capture technologies have enabled us to investigate the three-dimensional genome organization at an unprecedented resolution. However, high-resolution maps of chromatin interactions require costly, extremely deep sequencing and have been achieved for only a small number of cell lines. Without sufficient sequencing depth, the observed chromatin interaction maps are very sparse and noisy, which imposes great statistical and computational challenges. In this talk, I will present our recent work on enhancing the resolution of chromatin interaction maps via a generative adversarial network framework.

Spatial Transcriptomics Analysis

Guo-Cheng Yuan
Icahn School of Medicine at Mount Sinai
guo-cheng.yuan@mssm.edu

In a higher-eukaryotes organism, each organ contains multiple cell types in a highly organized manner. A fundamental biological question is how cells interact with each other to coordinate the function of the entire organ. The recent development of spatial transcriptomic technologies has provided an unprecedented opportunity to

systematically investigate the structural organization at single-cell resolution, yet it has also presented a number of new challenges for data analysis and interpretation. In the past few years, our lab has developed a number of methods (HMRF, binSpect, spatialDWLS, ICG) and built a generally applicable, comprehensive pipeline (Giotto) for spatial transcriptomic analysis. In this talk, I will give an introduction to these tools and applications.

MUNIn: A statistical framework for identifying long-range chromatin interactions from multiple samples

♦*Yuchen Yang*[1]*, Weifang Liu*[1]*, Yun Li*[1] *and Ming Hu*[2]
[1]University of North Carolina at Chapel Hill
[2]Cleveland Clinic Foundation
yyuchen@email.unc.edu

Chromatin spatial organization (interactome) plays a critical role in genome function. Deep understanding of chromatin interactome can shed insights into transcriptional regulation mechanisms and human disease pathology. One essential task in the analysis of chromatin interactomic data is to identify long-range chromatin interactions. Existing approaches, such as HiCCUPS, FitHiC/FitHiC2, and FastHiC, are all designed for analyzing individual cell types or samples. None of them accounts for unbalanced sequencing depths and heterogeneity among multiple cell types or samples in a unified statistical framework. To fill in the gap, we have developed a novel sta- tistical framework MUNIn (multiple-sample unifying long-range chromatin-interaction detector) for identifying long-range chromatin interactions from multiple samples. MUNIn adopts a hierarchical hidden Markov random field (H-HMRF) model, in which the status (peak or background) of each interacting chromatin loci pair depends not only on the status of loci pairs in its neighborhood region but also on the status of the same loci pair in other samples. To benchmark the performance of MUNIn, we performed comprehensive simulation studies and real data analysis and showed that MUNIn can achieve much lower false-positive rates for detecting sample-spe- cific interactions (33.1%-36.2%), and much enhanced statistical power for detecting shared peaks (up to 74.3%), compared to uni-sam- ple analysis. Our data demonstrated that MUNIn is a useful tool for the integrative analysis of interactomic data from multiple samples.

Session 71: Structural Inference of Time Series Data

Multiple change point detection under serial dependence

♦*Haeran Cho*[1] *and Piotr Fryzlewicz*[2]
[1]University of Bristol
[2]London School of Economics
haeran.cho@bristol.ac.uk

We propose a methodology for detecting multiple change points in the mean of an otherwise stationary, autocorrelated, linear time series. It combines solution path generation well-suited to separating shifts in the mean from fluctuations due to serial correlations, and an information criterion-based model selection strategy which simultaneously estimates the dependence structure and the number of change points without performing the difficult task of estimating the level of the noise as quantified e.g. by the long-run variance. We provide modular investigation into their theoretical properties and show that the combined methodology achieves consistency in estimating both the total number and the locations of the change points. Its good performance is demonstrated via extensive simulation studies, and we further illustrate its usefulness by applying the methodology to London air quality data.

Modeling the COVID-19 infection trajectory: a piecewise linear quantile trend model

♦*Feiyu Jiang*[1]*, Zifeng Zhao*[2] *and Xiaofeng Shao*[3]
[1]Fudan University
[2]University of Notre Dame
[3]University of Illinois at Urbana Champaign
jiangfy@fudan.edu.cn

We propose a piecewise linear quantile trend model to analyze the trajectory of the COVID-19 dailynew cases (i.e. the infection curve) simultaneously across multiple quantiles. The model is intuitive,interpretable and naturally captures the phase transitions of the epidemic growth rate via changepoints. Unlike the mean trend model and least squares estimation, our quantile-based approach isrobust to outliers, captures heteroscedasticity (commonly exhibited by COVID-19 infection curves)and automatically delivers both point and interval forecasts with minimal assumptions. Building on aself-normalized (SN) test statistic, a novel segmentation algorithm is proposed for multiple change-pointestimation. Theoretical guarantees such as segmentation consistency are established under mild andverifiable assumptions. Using the proposed method, we analyze the COVID-19 infection curves in 35major countries and discover interesting patterns with potentially relevant implications for effectivenessof the pandemic responses by different countries. A simple change-adaptive two-stage forecastingscheme is further designed to generate short-term prediction of COVID-19 cumulative new cases andis shown to deliver accurate forecast valuable to public health decision-making.

Testing long-range dependence for locally stationary time series nonparametric regression

♦*Lujia Bai and Weichi Wu*
Tsinghua University
blj20@mails.tsinghua.edu.cn

We study several KPSS-type tests for long memory in varying coefficient regression models. The responses and covariates considered are allowed to be locally stationary time series. We obtain the limiting distribution of the test statistics under the null hypothesis, local alternatives as well as the fixed alternative. We also provide a theoretically justified bootstrap approach for the implementation. The effectiveness of the tests is demonstrated by a simulation study and the analysis of real data examples.

A Composite Likelihood-based Approach for Change-point Detection in Spatio-temporal Process

Ting Fung Ma[1]*, Wai Leong Ng*[2]*,* ♦*Chun Yip Yau*[3] *and Zifeng Zhao*[4]
[1]University of Wisconsin, Madison
[2]Hang Seng University of Hong Kong
[3]Chinese University of Hong Kong
[4]University of Notre Dame
cyyau@sta.cuhk.edu.hk

This paper develops a unified, accurate and computationally efficient method for change-point inference in non-stationary spatio-temporal processes. By modeling a non-stationary spatio-temporal process as a piecewise stationary spatio-temporal process, we consider simultaneous estimation of the number and locations of change-points, and model parameters in each segment. A composite likelihood-based criterion is developed for change-point and parameters estimation. Under the framework of increasing domain asymptotics, asymptotic theories including consistency and distribution of the estimators are derived under mild conditions. In contrast to classical results in fixed dimensional time series that the asymptotic error of change-point estimator is $O_p(1)$, exact recovery of true change-points is guaranteed in the spatio-temporal set-

ting. More surprisingly, the consistency of change-point estimation can be achieved without any penalty term in the criterion function. Moreover, under an alternative asymptotic framework in which the time domain is increasing while the spatial sampling domain is fixed (i.e. the infill asymptotics), consistency of the number and locations of change-points is also established. A computational efficient pruned dynamic programming algorithm is developed for the challenging criterion optimization problem. Simulation studies and an application to U.S. precipitation data are provided to demonstrate the effectiveness and practicality of the proposed method.

Session 72: Advances in Forensic Statistics

Homogeneity test for ordinal ROC regression and application to facial recognition

♦*Ngoc-Ty Nguyen and Larry Tang*
National Center of Forensic Science, University of Central Florida
ty.nguyen@knights.ucf.edu

In facial recognition, ordinal scores are given by facial examiners to show confidence about whether persons in two images are the same or two different ones. Those scores can be analyzed by ROC curve to evaluate accuracy. In this talk, we propose a homogeneity test to compare performance of facial examiners. Asymptotic properties of estimated ROC curves and their corresponding AUCs within ordinal regression framework are derived as well. Moreover, behavior of difference in ROC cures (and AUCs) among examiners are investigated in detail. Confidence intervals of difference in AUCs and confidence bands of difference in ROC curves are built up to support for performance comparison purpose. Simulations are conducted on data where scores are assumed to come from binormal distribution and both categorical and continuous covariates are involved. Finally, we apply our procedure to facial recognition data to compare accuracy performance among image examiners.

The Effect of Latent Structures on Forensic Values of Evidence

Dylan Borchet, Andrew Simpson, Semhar Michael and ♦*Christopher Saunders*
South Dakota State University
christopher.saunders@sdstate.edu

During the various efforts associated with the National Institute of Standards and Technology it became clear that a process known as pre-screening was being performed prior to the assessment of evidential value of forensic evidence. Prescreening is a process by which the forensic examiner determines which control samples have sufficient information to support a reasonable comparison between the control samples and the samples of interest (in the computation of a likelihood ratio). In this work we are exploring the effect of the pre-screening process on the forensic likelihood ratio. We will discuss both the benefits and problems associated with presenting a likelihood ratio for source identification problems following a prescreening process.

Constructing Coherent Score-Based Likelihood Ratios that Account for Rarity

♦*Danica Ommen*[1] *and Nate Garton*[2]
[1]Iowa State University/CSAFE
[2]Commonwealth Computer Research, Inc.
dmommen@iastate.edu

Score-based likelihood ratios (SLRs) are the most practical alternative to feature-based likelihood ratios for the evaluation of the strength of forensic evidence. The construction of effective general score functions, however, has received little attention. Many scores are measures of dissimilarity between two pieces of evidence. However, it is not always obvious which two pieces of evidence should be compared. This leads to applications of score-based likelihood ratios that suffer from incoherence, e.g. when you change the order in which the hypotheses are considered and the resulting SLR value is different from what's expected. We will argue that this legitimate problem with SLRs should not be characterized as a lack of coherence, but rather a subtlety relating to the choice of an appropriate score function. Specifically, we will show that the standard argument as to why SLRs are incoherent can be understood as the comparison of two SLRs based on different score functions. This line of thought then leads to natural questions about how to construct scores even in the presence of an agreed upon dissimilarity metric. For example, another common criticism of SLR approaches is that they do not account for the rarity of the features in a relevant background population. Towards this end, we consider building scores as an aggregation of many dissimilarity metrics and discuss potential relationships between these approaches and rarity. Furthermore, we demonstrate that the resulting SLRs are both coherent and superior to standard scores via simulations.

Approaches to Likelihood Ratio Estimation for Forensic Evidence Interpretation

♦*He Qi and Martin Slawski*
George Mason University
hqi4@gmu.edu

Density ratios (aka score-based likelihood ratios) play in an important role in the interpretation of forensic evidence. In this talk, we will compare two different estimation methods based on kernel density estimation, the ratio of density estimators and a rank-invariant transformation-based density estimator. It is shown that the latter can achieve significantly better performance. We then discuss an extension of this approach to account for potential covariates of interest (subjects' demographics, forensic examiner, measurement characteristics, etc.) via a semiparametric location-scale model.

Session 73: Advances in selective and simultaneous inference

Post-selection inference: some answers and some questions

Arun KUCHIBHOTLA
Carnegie Mellon University
arunku@stat.cmu.edu

Data exploration is an important part of practical data analysis. This is also an important reason for the reproducibility crisis. In recent years numerous solutions for the problem of inference after data exploration (post-selection inference). In this talk, I will discuss simultaneous inference solutions to the post-selection problem with variable selection as well as variable transformation. I will also discuss some open questions.

Valid Inference After Hierarchical Clustering

♦*Lucy Gao*[1]*, Jacob Bien*[2] *and Daniela Witten*[3]
[1]University of Waterloo
[2]University of Southern California
[3]University of Washington
lucy.gao@uwaterloo.ca

As datasets continue to grow in size, in many settings the focus of data collection has shifted away from testing pre-specified hypotheses, and towards hypothesis generation. Researchers are often interested in performing an exploratory data analysis in order to generate hypotheses, and then testing those hypotheses on the same data; I

will refer to this as 'double dipping'. Unfortunately, double dipping can lead to highly-inflated Type 1 errors. In this talk, I will consider the special case of hierarchical clustering. First, I will show that sample-splitting does not solve the 'double dipping' problem for clustering. Then, I will propose a test for a difference in means between estimated clusters that accounts for the cluster estimation process, using a selective inference framework. I will also show an application of this approach to single-cell RNA-sequencing data. This is joint work with Jacob Bien (University of Southern California) and Daniela Witten (University of Washington).

Selective peak effect size inference

♦*Samuel Davenport*[1] *and Thomas Nichols*[2]
[1]University of California, San Diego
[2]University of Oxford
samuel.davenport@stats.ox.ac.uk

The spatial signals in neuroimaging mass univariate analyses can be characterized in a number of ways, but one widely used approach is peak inference: the identification of peaks in the image. Peak locations and magnitudes provide a useful summary of activation and are routinely reported, however, the magnitudes reflect selection bias as these points have both survived a threshold and are local maxima. This is known as the winner's curse or double dipping. In this talk I will discuss how to correct for this using a selective inference resampling based approach that allows the use of all of the data to estimate peak locations and effect sizes. I will discuss how this can be used in order to estimate both the raw units change, as well as standardized effect size measured with Cohen's d and partial R^2, and will present a big data validation of our methods using the UK biobank. For further reference our paper is available here: https://www.sciencedirect.com/science/article/pii/S1053811919309668.

Session 74: Advanced statistical inference for complex data structures

Hypothesis tests for block Markov chains

♦*Can Le and Russell Okino*
University of California, Davis
canle@ucdavis.edu

Consider the Block Markov Model with a finite number of states partitioned into a fixed number of disjoint groups such that the transition probabilities at any state only depend on the state's group membership. Given some information of a path drawn from this model, we propose a few goodness-of-fit test statistics and study their properties. We show that these statistics converge to standard normal random variables as long as the length of the path grows faster than the number of states. Through simulations, we also show that these tests exhibit good power under natural alternatives.

Change Point Detection in Network Sequences

♦*Sharmodeep Bhattacharyya*[1]*, Shirshendu Chatterjee*[2]*, Shyamal De*[3] *and Soumendu Mukherjee*[3]
[1]Oregon State University
[2]City University of New York
[3]Indian Statistical Institute, Kolkata
sharmodeep.bhattacharyya@oregonstate.edu

We consider the problem of local change point detection in a sequence of network data. We propose a class of models, called Transitive Inhomogeneous Erdos-Renyi (TIER) models, which are generalizations of the Inhomogeneous Erdos-Renyi model with dependent edges in each network layer. We perform change-point detection using a sequence of networks generated from the TIER models. We use local subgraph statistics to estimate change points in node-specific localities of the networks. We also provide a theoretical analysis of the change point detection method for network sequences with each network layer generated from the TIER model and its variants. We validate our results using simulation studies too.

Partial Recovery for Top-k Ranking: Optimality of MLE and Sub-Optimality of Spectral Method

Anderson Ye Zhang
University of Pennsylvania
ayz@wharton.upenn.edu

Given partially observed pairwise comparison data generated by the Bradley-Terry-Luce (BTL) model, we study the problem of top-k ranking. That is, to optimally identify the set of top-k players. We derive the minimax rate with respect to a normalized Hamming loss. This provides the first result in the literature that characterizes the partial recovery error in terms of the proportion of mistakes for top-k ranking. We also derive the optimal signal-to-noise ratio condition for the exact recovery of the top-k set. The maximum likelihood estimator (MLE) is shown to achieve both optimal partial recovery and optimal exact recovery. On the other hand, we show another popular algorithm, the spectral method, is in general sub-optimal.

Session 75: Exploratory Functional Data Analysis

Functional outlier detection and taxonomy by sequential transformations

♦*Wenlin Dai*[1]*, Tomas Mrkvicka*[2]*, Ying Sun*[3] *and Marc G. Genton*[3]
[1]Renmin University of China
[2]University of South Bohemia
[3]King Abdullah University of Science and Technology
wenlin.dai@ruc.edu.cn

Functional data analysis can be seriously impaired by abnormal observations, which can be classified as either magnitude or shape outliers based on their way of deviating from the bulk of data. Identifying magnitude outliers is relatively easy, while detecting shape outliers is much more challenging. We propose turning the shape outliers into magnitude outliers through data transformation and detecting them using the functional boxplot. Besides easing the detection procedure, applying several transformations sequentially provides a reasonable taxonomy for the flagged outliers. A joint functional ranking, which consists of several transformations, is also defined here. Simulation studies are carried out to evaluate the performance of the proposed method using different functional depth notions. Interesting results are obtained in several practical applications.

Covariance function visualization using functional data analysis

♦*Huang Huang, Ying Sun and Marc Genton*
King Abdullah University of Science and Technology
huang.huang@kaust.edu.sa

The prevalence of multivariate space-time data collected from monitoring networks and satellites or generated from numerical models has brought much attention to multivariate spatio-temporal statistical models, where the covariance function plays a key role in modeling, inference, and prediction. For multivariate space-time data, understanding the spatio-temporal variability, within and across variables, is essential in employing a realistic covariance model. Meanwhile, the complexity of generic covariances often makes model fitting very challenging, and simplified covariance structures, including symmetry and separability, can reduce the model complexity

and facilitate the inference procedure. However, a careful examination of these properties is needed in real applications. In this work, we formally define these properties for multivariate spatio-temporal random fields and use functional data analysis techniques to visualize them, hence providing intuitive interpretations. We then propose a rigorous rank-based testing procedure to conclude whether the simplified properties of covariance are suitable for the underlying multivariate space-time data. The good performance of our method is illustrated through synthetic data, for which we know the true structure. We also investigate the covariance of bivariate wind speed, a key variable in renewable energy, over a coastal and an inland area in Saudi Arabia.

Sparse Functional Boxplots for Multivariate Curves

Zhuo Qu and ♦*Marc G. Genton*
KAUST
marc.genton@kaust.edu.sa

This talk introduces the sparse functional boxplot and the intensity sparse functional boxplot as practical exploratory tools that make visualization possible for both complete and sparse functional data. These visualization tools can be used either in the univariate or multivariate functional setting. The sparse functional boxplot, which is based on the functional boxplot, depicts sparseness characteristics in the envelope of the 50% central region, the median curve, and the outliers. The proportion of missingness at each time index within the central region is colored in gray. The intensity sparse functional boxplot displays the relative intensity of sparse points in the central region, revealing where data are more or less sparse. The two-stage functional boxplot, a derivation from the functional boxplot to better detect outliers, is also extended to its sparse form. Several depth proposals for sparse multivariate functional data are evaluated and outlier detection is tested in simulations under various data settings and sparseness scenarios. The practical applications of the sparse functional boxplot and intensity sparse functional boxplot are illustrated with two public health datasets.

Exploratory Functional Data Analysis

Hanlin Shang
Macquarie University
hanlin.shang@mq.edu.au

Discussant for an invited session titled "Exploratory Functional Data Analysis"

Session 76: New Frontiers in Precision Medicine

Learning Individualized Treatment Rules for Multiple-Domain Latent Outcomes

Yuan Chen[1]*, Yuanjia Wang*[2] *and* ♦*Donglin Zeng*[3]
[1]yc3281@cumc.columbia.edu
[2]yw2016@cumc.columbia.edu
[3]dzeng@email.unc.edu
dzeng@email.unc.edu

For many mental disorders, latent mental status from multiple domain psychological or clinical symptoms perform as a better characterization of the underlying disorder status than a simple summary score of the symptoms, and they also serve as a more reliable and representative features to differentiate treatment responses. We provide a new paradigm for learning optimal individualized treatment rules (ITRs) for patients' latent mental status. We first learn the multi-domain latent states from the observed symptoms at baseline under a restricted Boltzmann machine (RBM) model. We then optimize a value function for the latent states using a weighted large margin classifier to estimate the optimal ITRs. We derive the convergence rate of the resulting value when applied to a future population. Simulation studies are conducted to test the performance of the proposed method. Finally, we apply the developed method to a real world study for patients with major depression, and identify patient subgroups informative for treatment recommendations.

Estimation and inference on individualized treatment rule in observational data.

Yingqi Zhao
Fred Hutchinson Cancer Research Center
yqzhao@fredhutch.org

With the increasing adoption of electronic health records, there is an increasing interest in developing individualized treatment rules (ITRs), which recommend treatments according to patients' characteristics, from large observational data. In this talk, I will first introduce an improved doubly robust estimator of the optimal ITRs. The method enjoys two key properties. First, it is doubly robust, meaning that the proposed estimator is consistent when either the propensity score or the outcome model is correct. Second, it achieves the smallest variance among the class of doubly robust estimators when the propensity score model is correctly specified, regardless of the specification of the outcome model. I will then introduce a penalized doubly robust method to estimate the optimal ITRs from high-dimensional observational data, along with a split-and-pooled de-correlated score to construct hypothesis tests and confidence intervals. This method utilizes the data splitting to conquer the slow convergence rate of nuisance parameter estimations, such as nonparametric methods for outcome regression or propensity models. Simulation and real data analysis are conducted to demonstrate the superiority of the proposed methods.

Experimental Design and Causal Inference Methods For Micro-Randomized Trials: A Framework for Developing Mobile Health Interventions

♦*Tianchen Qian*[1]*, Predrag Klasnja*[2] *and Susan Murphy*[3]
[1]University of California, Irvine
[2]University of Michigan
[3]Harvard University
t.qian@uci.edu

Mobile devices along with wearable sensors facilitate our ability to deliver supportive behavioral interventions to individuals anytime and anywhere. These interventions are being developed and employed across a variety of health fields, including to improve medication adherence, to encourage physical activity and healthier eating, and to support recovery in addictions. Most mobile health interventions involve notifications such as reminders or push notifications, delivered as an individual goes about their everyday life. Yet repeated notifications can lead to disengagement. To reduce disengagement and improve effectiveness, it is critical to only deliver notifications when they are most likely to be effective, which can be different for different individuals. Thus, critical questions in the optimization of mobile health interventions include not only, "Are the notifications useful (on average)?" but also "When and in which contexts is it most useful to deliver notifications to the individual?" This question concerns time-varying dynamic moderation by the context (location, stress, time of day, etc.) of the effectiveness of the mobile health interventions on an individual's behavior. In this talk, we discuss the micro-randomized trial design and associated causal inference methods for use in assessing such effect moderation. We illustrate the ideas with the micro-randomized trials across a variety of fields.

Causal Inference on Non-linear Spaces: Distribution Functions and Beyond

Zhenhua Lin[1], ♦*Dehan Kong*[2] *and Linbo Wang*[2]
[1]National University of Singapore
[2]University of Toronto
kongdehan@utstat.toronto.edu

Understanding causal relationships is one of the most important goals of modern science. So far, the causal inference literature has focused almost exclusively on outcomes coming from a linear space, most commonly the Euclidean space. However, it is increasingly common that complex datasets collected through electronic sources, such as wearable devices and medical imaging, cannot be represented as data points from linear spaces. In this paper, we present a formal definition of causal effects for outcomes from non-linear spaces, with a focus on the Wasserstein space of cumulative distribution functions. We develop doubly robust estimators and associated asymptotic theory for these causal effects. Our framework extends to outcomes from certain Riemannian manifolds. As an illustration, we use our framework to quantify the causal effect of marriage on physical activity patterns using wearable device data collected through the National Health and Nutrition Examination Survey.

Session 77: Efficient estimation methods for clinical trials

everaging auxiliary covariates to improve efficiency of inferences: a general framework and practical considerations

♦*Min Zhang*[1] *and Baqun Zhang*[2]
[1]University of Michigan
[2]Shanghai University of Finance and Economics
mzhangst@umich.edu

Auxiliary covariates are routinely collected in randomized clinical trials. Although it has been long recognized in statistical literature that incorporating covariates can improve the efficiency of inferences and reduce chance imbalance, covariates adjustment has not been used as often as it should be in the primary analysis of clinical trials in practice. This is partly due to that many practitioners remain skeptical of its usefulness or have other concerns regarding model misspecifications. We try to address these concerns and discuss a general framework that allows one to adjust covariates robustly. That is, this framework can guarantee improvement in efficiency, if covariates are predictive of outcomes, and valid inferences regardless of whether the specified models for covariates are correct or not. Therefore, it can eliminate the concern over model misspecification. We further discuss practical considerations in terms of covariate adjustment, focusing on finite sample effects, stratified randomization, and what variables to adjust. We attempt to address the question of how and when to use covariate adjustment for randomized clinical trials in practice.

Covariate adjustment in randomized studies with ordinal and time-to-event endpoints

Ivan Diaz
Weill Cornell Medicine
ild2005@med.cornell.edu

We present new estimators for ordinal and time-to-event outcomes in randomized trials that do not rely on proportional odds/hazard assumptions. The proposed estimators leverage prognostic baseline variables to obtain equal or better asymptotic precision compared to traditional estimators. The proposed estimators have the following features: (i) they are interpretable under violations of the proportional odds/hazards assumption; (ii) they are consistent and at least as precise as the unadjusted estimators; (iii) for time-to-event outcomes, they remain consistent under violations of independent censoring (unlike the Kaplan-Meier estimator) when either the censoring or survival distributions, conditional on covariates, are estimated consistently; and (iv) they achieve the nonparametric efficiency bound when both of these distributions are consistently estimated. We illustrate the performance of our method using simulations based on resampling data from various completed clinical trials and from hospitalized COVID patient data. We found substantial precision gains from using covariate adjustment–equivalent to 9-21% reductions in the required sample size to achieve a desired power–for a variety of estimands (targets of inference) when the trial sample size was at least 200

Model-Robust Inference for Clinical Trials that Improve Precision by Stratified Randomization and Covariate Adjustment

♦*Bingkai Wang*[1], *Ryoko Susukida*[2], *Ramin Mojtabai*[2], *Masoumeh Amin-Esmaeili*[2] *and Michael Rosenblum*[3]
[1]The statistics and data science department of the Wharton School, University of Pennsylvania
[2]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health
[3]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
bingkai.w@gmail.com

Two commonly used methods for improving precision and power in clinical trials are stratified randomization and covariate adjustment. However, many trials do not fully capitalize on the combined precision gains from these two methods, which can lead to wasted resources in terms of sample size and trial duration. We derive consistency and asymptotic normality of model-robust estimators that combine these two methods, and show that these estimators can lead to substantial gains in precision and power. Our theorems cover a class of estimators that handle continuous, binary, and time-to-event outcomes; missing outcomes under the missing at random assumption are handled as well. For each estimator, we give a formula for a consistent variance estimator that is model-robust and that fully captures variance reductions from stratified randomization and covariate adjustment. Also, we give the first proof (to the best of our knowledge) of consistency and asymptotic normality of the Kaplan-Meier estimator under stratified randomization, and we derive its asymptotic variance. The above results also hold for the biased-coin covariate-adaptive design. We demonstrate our results using data from three trials of substance use disorder treatments, where the variance reduction due to stratified randomization and covariate adjustment ranges from 1% to 36%.

Data-Adaptive Efficient Estimation Strategies for Biomarker Studies Embedded in Randomized Trials

Wei Zhang[1], ♦*Zhiwei Zhang*[2], *James Troendle*[2] *and Aiyi Liu*[2]
[1]Chinese Academy of Sciences
[2]National Institutes of Health
zhiwei.zhang@nih.gov

Predictive and prognostic biomarkers are increasingly important in clinical research and practice. Biomarker studies are frequently embedded in randomized clinical trials with biospecimens collected at baseline and assayed for biomarkers, either in real time or retrospectively. In this work, we propose efficient estimation strategies for two study settings in terms of biomarker ascertainment: the complete-data setting in which the biomarker is measured for all subjects in the trial, and a two-phase sampling design in which the biomarker is measured retrospectively for a random subsample of

subjects selected in an outcome-dependent fashion. In both settings, efficient estimating functions are characterized using semiparametric theory and approximated using data-adaptive machine learning methods, leading to estimators that are consistent, asymptotically normal, and (approximately) efficient under general conditions. The proposed methods are evaluated in simulation studies and applied to real data from a cancer trial.

Session 78: Recent developments in regression methods for bio-medical studies

Multicategory Outcome Weighted Learning for Dynamic Treatment Regimes

♦*Wenqing He and Junwei Shen*
University of Western Ontario
whe@stats.uwo.ca

Dynamic treatment regimes are sequential decision rules dictating how to individualize treatments to patients based on evolving treatments and covariate history. The outcome weighted learning methods are introduced to obtain the optimal treatment rules. In this talk, we extend the outcome weighted learning from two-treatments to multi-treatments and allows for negative treatment outcome. We show that under two different sets of assumptions, the Fisher consistency can be maintained. Simulation studies and real application to data from Sequential Treatment Alternatives to Relieve Depression (STAR*D) clinical trial are conducted to illustrate the proposed methods.

A Semiparametric Isotonic Regression Model for Skewed Distributions

♦*Chenguang Wang*[1]*, Ao Yuan*[2]*, Leslie Cope*[1] *and Jing Qin*[3]
[1]Johns Hopkins University
[2]Georgetown University
[3]National Institute of Allergy and Infectious Diseases
cwang68@jhmi.edu

In this work, we propose a semiparametric regression model that is built upon an isotonic regression model with the assumption that the random error follows a skewed distribution. We develop an EM algorithm for obtaining the maximum likelihood estimates of the model parameters, examine the asymptotic properties of the estimators, conduct simulation studies to explore the performance of the proposed model, and apply the method to evaluate the DNA-RNA-protein relationship and identify genes that are key factors in tumor progression.

Order-Constrained ROC Regression with Application to Facial Recognition

Larry Tang
University of Central Florida
tangstat@gmail.com

We consider modeling of ROC curves using both the order constraint and covariates associated with each score given that the latter (e.g., demographic characteristics of the underlying subjects) often have a substantial impact on discriminative accuracy. The proposed method is based on the indirect ROC regression approach using a location-scale model, and quadratic optimization is used to implement the order constraint. The statistical properties of the proposed order-constrained least squares estimator are studied. Based on the theoretical results developed herein, we deduce that the proposed estimator can achieve substantial reductions in mean squared error relative to its unconstrained counterpart.

Inferring random change point with segmented non-linear mixed effect models

Hongbin Zhang
City University of New York
hongbin.zhang@sph.cuny.edu

The primary goal of public health efforts which aim to control HIV epidemics in the United States and around the world is to diagnose and treat people with HIV infection as soon as possible after seroconversion. The timing of the first antiretroviral therapy (ART) treatment after HIV diagnosis is therefore an important population-level indicator that can be used to measure the effectiveness of HIV care programs and policies at local and national levels. In this paper, we present a random change-point model based method to estimation ART initiation time utilizing routinely reported individual level viral loads. To deal with the left-censoring occurred to viral loads and the nonlinear trajectory, we formulate a flexible segmented nonlinear mixed effects model and propose a Stochastic version of EM (StEM) algorithm, coupled with Gibbs sampler for the inference. We apply the method to an HIV surveillance data to estimate ART initiation after diagnosis and to gain additional insights of the viral load dynamics. Simulation studies are also performed to evaluate the properties of the proposed method.

Session 79: Recent Advance of Design of Experiments in Online Experimentation and Personalized Medicine

Novelty and Primacy: A Long-Term Estimator for Online Experiments

♦*Sol Sadeghi, Somit Gupta, Stefan Gramatovici, Jiannan Lu, Hao Ai and Ruhan Zhang*
Microsoft
Soheil.Sadeghi@microsoft.com

Online experiments are the gold standard for evaluating impact on user experience and accelerating innovation in software. However, since experiments are typically limited in duration, observed treatment effects are not always permanently stable, sometimes revealing increasing or decreasing patterns over time. There are multiple causes for a treatment effect to change over time. In this paper, we focus on a particular cause, user-learning, which is primarily associated with novelty or primacy. Novelty describes the desire to use new technology that tends to diminish over time. Primacy describes the growing engagement with technology as a result of adoption of the innovation. User-learning estimation is critical because it holds experimentation responsible for trustworthiness, empowers organizations to make better decisions by providing a long-term view of expected impact, and prevents user dissatisfaction. In this paper, we propose an observational approach, based on difference-in-differences technique to estimate user-learning at scale. We use this approach to test and estimate user-learning in many experiments at Microsoft. We compare our approach with the existing experimental method to show its benefits in terms of ease of use and higher statistical power, and to discuss its limitation in presence of other forms of treatment interaction with time.

Robust sequential design for piecewise-stationary multi-armed bandit problem in the presence of outliers

Yaping Wang[1]*, Zhicheng Peng*[1]*, Riquan Zhang*[1] *and* ♦*Qian Xiao*[2]
[1]East China Normal University
[2]University of Georgia
QIAN.XIAO@uga.edu

The multi-armed bandit (MAB) problem studies the sequential decision making in the presence of uncertainty and partial feedback

on rewards. Its name comes from imagining a gambler at a row of slot machines who needs to decide the best strategy on the number of times as well as the orders to play each machine. It is a classic reinforcement learning problem which is fundamental to many online learning problems. In many practical applications of the MAB, the reward distributions may change at unknown time steps and the outliers (extreme rewards) often exist. Current sequential design strategies may struggle in such cases, as they tend to infer additional change points to fit the outliers. In this paper, we propose a robust change-detection upper confidence bound (RCD-UCB) algorithm which can distinguish the real change points from the outliers in piecewise-stationary MAB settings. We show that the proposed RCD-UCB algorithm can achieve a nearly optimal regret bound on the order of $O(\sqrt{SKT\log T})$, where T is the number of time steps, K is the number of arms and S is the number of stationary segments. We demonstrate its superior performance compared to some state-of-the-art algorithms in both simulation experiments and real data analysis.

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

♦*Qiong Zhang, Amin Khademi and Yongjia Song*
Clemson University
qiongz@clemson.edu

In this talk, I will discuss the optimal design of two-armed clinical trials to maximize the accuracy of parameter estimation in a statistical model, where the interaction between patient covariates and treatment are explicitly incorporated to enable precision medication decisions. Such a modeling extension leads to significant complexities for the produced optimization problems because they include optimization over design and covariates concurrently. We take a min-max optimization model and minimize (over design) the maximum (over population) variance of the estimated interaction effect between treatment and patient covariates. This results in a min-max bi-level mixed integer nonlinear programming problem, which is notably challenging to solve. To address this challenge, we introduce a surrogate optimization model by approximating the objective function and propose a lower bound for the inner optimization problem and solve the outer optimization problem over the lower bound. We test our proposed algorithms with synthetic and real-world data sets and compare them with standard (re-)randomization methods.

Recent Advance in DoE Driven by New Applications and/or Algorithms

Lulu Kang
Illinoise Institue of Techonogy
lkang2@iit.edu

Design of Experiments, although a classic topic in statistics, has seen significant advance in recent years, driven by new applications and new algorithms. Three such examples are introduced in this session. In this talk, I give more such examples and discuss what could be some future directions of the DoE research field.

Session 80: Statistical Modeling for the COVID-19 Pandemic

A Spatiotemporal Epidemiological Prediction Model to Inform County-level COVID-19 Risk in the USA

Yiwang Zhou, Lili Wang and ♦*Peter Song*
University of Michigan
pxsong@umich.edu

We develop a health information system that provides micro COVID-19 infection risk predictions in a similar way to the weather forecast. Generalizing the von Neumann's cellular automata with the stochastic infectious disease models, this forecast system integrates multiple sources of information in the risk prediction, including serological test surveys, mobile device data and personal mobility scores. This prediction model is tuned by minimizing the prediction error of one-day ahead projection of infection prevalence. We illustrate the proposed method by the county-level COVID-19 prevalence projection over 3,109 counties in the continental US. In comparison to the conventional temporal risk prediction models, our model informs high-resolution community-level risk projection, which is useful for tailored decision-making on business reopenings and medical resource allocation.

Robust estimation of SARS-CoV-2 epidemic in US counties

Hanmo Li and ♦*Mengyang Gu*
University Of California, Santa Barbara
mengyang@pstat.ucsb.edu

The COVID-19 outbreak is asynchronous in US counties. Mitigating the COVID-19 transmission requires not only the state and federal level order of protective measures such as social distancing and testing, but also public awareness of time-dependent risk and reactions at county and community levels. We propose a robust approach to estimate the heterogeneous progression of SARS-CoV-2 at all US counties having no less than 2 COVID-19 associated deaths, and we use the daily probability of contracting (PoC) SARS-CoV-2 for a susceptible individual to quantify the risk of SARS-CoV-2 transmission in a community. We found that shortening by 5% of the infectious period of SARS-CoV-2 can reduce around 39% (or 78 K, 95% CI: [66 K , 89 K]) of the COVID-19 associated deaths in the US as of 20 September 2020. Our findings also indicate that reducing infection and deaths by a shortened infectious period is more pronounced for areas with the effective reproduction number close to 1, suggesting that testing should be used along with other mitigation measures, such as social distancing and facial mask-wearing, to reduce the transmission rate. Our deliverable includes a dynamic county-level map for local officials to determine optimal policy responses and for the public to better understand the risk of contracting SARS-CoV-2 on each day.

SaucIR: a Migration-based Model for Forecasting Confirmed Cases of the COVID-19 Pandemic

Xinyu Wang[1]*, Lu Yang*[1]*, Hong Zhang*[1]*, Zhouwang Yang*[1] *and* ♦*Catherine Liu*[2]
[1]University of Science and Technology of China
[2]The Hong Kong Polytechnic University
catherine.chunling.liu@polyu.edu.hk

The unprecedented coronavirus disease 2019 (COVID-19) pandemic is still a worldwide threat to human life since its invasion into daily lives of the public in the first several months of 2020. Predicting the size of confirmed cases is important for countries and communities to make proper prevention and control policies so as to effectively curb the spread of COVID-19. Different from the 2003 SARS epidemic and the worldwide 2009 H1N1 influenza pandemic, COVID-19 has unique epidemiological characteristics in its infectious and recovered compartments. This drives us to formulate a new infectious dynamic model for forecasting the COVID-19 pandemic within the human mobility network, named the SaucIR-model in the sense that the new compartmental model extends the benchmark SIR model by dividing the flow of people in the infected state into asymptomatic, pathologically infected but unconfirmed,

and confirmed. Furthermore, we employ dynamic modeling of population flow in the model in order that spatial effects can be incorporated effectively. We forecast the spread of accumulated confirmed cases in some provinces of mainland China and other countries that experienced severe infection during the time period from late February to early May 2020. The novelty of incorporating the geographic spread of the pandemic leads to a surprisingly good agreement with published confirmed case reports. The numerical analysis validates the high degree of predictability of our proposed SaucIR model compared to existing resemblance.

Modelling biological change in COVID-19 patient using nonparametric mixed effect mixture model

Xiaoran Ma
University of California Santa Barbara
xiaoran_ma@umail.ucsb.edu

Some, but not all, COVID-19 patients had changes in biological variables such as temperature and oxygen sat-uration days before symptoms occur. We propose a flexible nonparametric mixed-effects mixture model (NMEM) that simultaneously identifies risk factors and classify patients with biological change. We model the latent biologicalchange probability using a logistic regression model and trajectories in each latent class using splines. We apply theEM algorithm and penalized likelihood to estimate all parameters and mean functions. Simulation studies indicatethe proposed method performs well. We apply the NMEM model to investigate changes in temperature and oxygensaturation in COVID-19 patients receiving hemodialysis.

Session 81: Incorporating RWE in regulatory submission

Indirect comparisons using real world data: confounding and population adjustment for time-to event endpoints

♦*Jixian Wang, Hongtao Zhang and Ram Tiwari*
BMS
Jixian.Wang@bms.com

There has been increasing use of real world data (RWD) as an external control arm or a comparator in standard clinical practice for internal clinical trials in regulatory submission and health technology assessment. Since often the trial and RWD populations are rather different in terms of key prognostic factors, proper adjustment for these confounding factors is a key step to make subjects in the trial and RWD comparable for valid treatment effect estimation. Although several approaches, such as propensity score matching/weighting and calibration weighting are available, using them for the time-to-event (TTE) endpoint comparison presents extra challenges due to the nature of TTE. We present a critical review of existing approaches with emphases on two directions: 1) the issue, assumptions for indirect comparison using hazard ratio as the estimand and its alternatives; 2) indirect comparison for TTE with a cure fraction. The abovementioned approaches are compared for TTE endpoints in some selected scenarios to illustrate the importance of careful selection of approaches and implementation. The simulation results together with practical advices will be presented.

Use of external control for single arm registrational trial: a case study in NSCLC

♦*Jianchang Lin and Qing Li*
Takeda
jianchang.lin@takeda.com

Most of the time, randomized clinical trials are preferred as the basis of an NDA/BLA submission. Nevertheless, there have been cases of trials that adopted single-arm designs that were successfully approved in FDA expedited programs for drugs and biologics treating serious conditions and fulfilling unmet medical needs (FDA 2014a). However, it could be challenging to construct an objective reference level to compare with a single-arm study. In this situation, it would be helpful to employ an external control as a comparison arm instead of presenting just the single experimental arm. These external control arms could be constructed from historical clinical trials, natural history studies, patient registry data, and other types of RWD. In this presentation, we will exemplify a case study in Non-Small Cell Lung Cancer (NSCLC) that used a single-arm trial design with external control for an NDA to illustrate the challenges and opportunities of using external control in clinical development. Additional considerations, including the selection of patient populations, endpoints, baseline covariates, propensity score methods, sensitivity analysis, and practical implementation flow in clinical development will be discussed.

Real-World Data in Regulatory Science

Diqiong Xie
Seagen
dxie@seagen.com

Real-world data include data from a wide range of sources, such as hospital charts, insurance claims, observation studies and clinical trials. RWD provides tremendous information of how patients are treated and affected by various therapies in real world. In this presentation, the author will discuss the types of RWD, study designs and an example of a comparative study using RWD.

Use of Real-World Evidence (RWE) in HTA Decision Making

Julia Ma
AbbVie
julia.y.ma@abbvie.com

Although randomized clinical trials (RCTs) have been perceived as the gold standard to demonstrate efficacy and safety, RWE gathered outside of RCTs now plays an increasingly important role in health technology assessment (HTA) submissions. RWE from sources such as observational studies, registry data have been submitted to HTA bodies to identify treatment patterns, characterize patients, and to provide supportive evidence for the economic evaluations. The focus of this talk is on the utility and the strengths of well-developed RWE in HTA decision making. While most HTA bodies become more open to RWE, the practice varies among the agencies and guidance development is also at different stages. We will also discuss the recent trend of incorporating RWE in HTA submissions, focusing on several leading HTA bodies.

Session 82: Flexible and efficient Bayesian methods for complex data modeling

New Directions in Bayesian Shrinkage for Structured Data

Jyotishka Datta
Virginia Tech
jyotishka@vt.edu

Sparse signal recovery remains an important challenge in large scale data analysis and global-local (G-L) shrinkage priors have undergone an explosive development in the last decade in both theory and methodology. These developments have established the G-L priors as the state-of-the-art Bayesian tool for sparse signal recovery as well as default non-linear problems. In the first half of my talk, I will survey the recent advances in this area, focusing on optimality and performance of G-L priors for both continuous as well

as discrete data. In the second half, I will discuss several recent developments, including designing a shrinkage prior to handle bi-level sparsity in regression and handling sparse compositional data, routinely observed in microbiomics. I will discuss the methodological challenges associated with each of these problems, and propose to address this gap by using new prior distributions, specially designed to enable handling structured data.

A Bayesian Selection Model for Correcting and Quantifying the Impact of Outcome Reporting Bias in Multivariate Meta-Analysis

♦*Ray Bai*[1]*, Lifeng Lin*[2]*, Mary Boland*[3] *and Yong Chen*[3]
[1]University of South Carolina
[2]Florida State University
[3]University of Pennsylvania
rbai@mailbox.sc.edu

Multivariate meta-analysis (MMA) is a powerful tool for jointly estimating multiple population treatment effects. However, the validity of results from MMA is potentially compromised by outcome reporting bias (ORB), or the tendency for studies to selectively report some outcomes while omitting others. Until recently, ORB has been understudied. Since ORB can lead to qualitative differences in the conclusions from MMA, it is crucial to both correct the estimates of effect sizes and quantify their uncertainty in the presence of ORB. With this goal, we develop a Bayesian selection model to correct for ORB in MMA. We further introduce a measure for quantifying the impact of ORB on the results from MMA. We evaluate our proposed approaches through simulations and a meta-evaluation of 782 bivariate meta-analyses from the Cochrane Database of Systematic Reviews. In addition, we apply our model to a real meta-analysis on the effects of interventions on quality of life and hospital readmission for heart failure patients. Under our model, the effect of intervention on hospital readmission is no longer statistically significant after we adjust for ORB. This case study demonstrates that failing to correct for ORB in MMA can lead to different conclusions in systematic research synthesis.

An approximate Bayesian approach to covariate dependent graphical modeling

Sutanoy Dasgupta, Prasenjit Ghosh, ♦*Debdeep Pati and Bani Mallick*
Texas A&M University
debdeep@tamu.edu

Gaussian graphical models typically assume a homogeneous structure across all subjects, which is often restrictive in applications. In this article, we propose a weighted pseudo-likelihood approach for graphical modeling which allows different subjects to have different graphical structures depending on extraneous covariates. The pseudo-likelihood approach replaces the joint distribution by a product of the conditional distributions of each variable. We cast the conditional distribution as a heteroscedastic regression problem, with covariate-dependent variance terms, to enable information borrowing directly from the data instead of a hierarchical framework. This allows independent graphical modelling for each subject, while retaining the benefits of a hierarchical Bayes model and being computationally tractable. An efficient embarrassingly parallel variational algorithm is developed to approximate the posterior and obtain estimates of the graphs. Using a fractional variational framework, we derive asymptotic risk bounds for the estimate in terms of a novel variant of the α-Renyi divergence. We theoretically demonstrate the advantages of information borrowing across covariates over independent modelling across covariates. We show the practical advantages of the approach through simulation studies and illustrate the dependence structure in protein expression levels on breast cancer patients using CNV information as covariates.

A Phase I-II Basket Trial Design to Optimize Dose-Schedule Regimes Based on Delayed Outcomes

Ruitao Lin
The University of Texas MD Anderson Cancer Center
RLin@mdanderson.org

I will present a Bayesian adaptive basket trial design to optimize the dose-schedule regimes of an experimental agent within disease subtypes, called "baskets", for phase I-II clinical trials based on late-onset efficacy and toxicity. To characterize the association among the baskets and regimes, a Bayesian hierarchical model is assumed that includes a heterogeneity parameter, adaptively updated during the trial, that quantifies information shared across baskets. To account for late-onset outcomes when doing sequential decision making, unobserved outcomes are treated as missing values and imputed by exploiting early biomarker and low-grade toxicity information. Elicited joint utilities of efficacy and toxicity are used for decision making. Patients are randomized adaptively to regimes while accounting for baskets, with randomization probabilities proportional to the posterior probability of achieving maximum utility. Simulations are presented to assess the design's robustness and ability to identify optimal dose-schedule regimes within disease subtypes, and to compare it to a simplified design that treats the subtypes independently.

Session 83: New advances in complex biomedical data analysis and the novel applications

Simple method for detecting the effect of exposure mixture

Xinhua Liu and ♦*Zhezhen Jin*
Columbia University
zj7@cumc.columbia.edu

In this talk, I presents a working model based approach for the study on the effect of exposure mixture on the continuous health outcomes. The approaches are useful for testing specific hypotheses including detecting overall effect and detecting unequal weights when the overall effect is evident. Under the null hypotheses, unbiased estimate of the parameter for overall effect is developed. The approaches are easy to implement with existing statistical software. Numerical examples will be presented for illustration.

Integrative Clustering Analysis with Application in Multi-Source Gene Expression Data

Liuqing Yang[1]*, Yunpeng Zhao*[2] *and* ♦*Qing Pan*[1]
[1]George Washington University
[2]Arizona State University
qpan@email.gwu.edu

In omics studies, different sources of information about the same set of genes are often available. When the group structure (e.g., gene pathways) within the genes are of interests, we combine the normal hierarchical model with the stochastic block model, through an integrative clustering framework, to model gene expression and gene networks jointly. The integrative framework provides higher accuracy in extensive simulation studies when one or both of the data sources contain noises or when different data sources provide complementary information. An empirical guideline in the choice between integrative versus separate clustering models is proposed. The integrative clustering method is illustrated on the mouse embryo single cell RNAseq and bulk cell microarray data, which iden-

tified not only the gene sets shared by both data sources but also the gene sets unique in one data source.

Peer-to-Peer Masking for Privacy-Preserving Medical Data Sharing

Hanzhi Gao[1]*, Dimitrios Melissourgos*[1]*, Shigang Chen*[1]*, Adam Ding*[2] *and* ♦*Samuel S. Wu*[1]
[1]University of Florida
[2]Northeastern University
gatorwu@gmail.com

A major hurdle in modern medical research is the hefty barrier to data sharing due to patient privacy concerns. One way to address this important problem is for the medical institutes to keep their raw data private while sharing only the masked data that are provably secure and statistically useful. The existing work has serious limitations in computational scalability, security loopholes, or impractical requirements. This paper proposes a new decentralized data perturbation protocol that is scalable, easy to deploy, and provably secure. It performs data obfuscation in a pure peer-to-peer fashion among the participating medical institutes through matrix transformations. The protocol does not rely on any trusted third parties, and it can resist against up to m-1 colluding participants trying to reveal the raw data of the sole victim participant, where m is number of participants. We show that the masked data supports the general linear model and contingency table analysis, which are widely used in the medical field, with the output from the masked data being exactly the same as that from the original data.

Event-Specific Win Ratios for Inference with Semi-Competing Risks Data

♦*Song Yang*[1]*, James Troendle*[2]*, Daewoo Pak*[3] *and Eric Leifer*[2]
[1]NIH NHBLI
[2]NIH NHLBI
[3]Yonsei University, South Korea
yangso@nhlbi.nih.gov

For semi-competing risks data involving a non-terminal event and a terminal event we derive the asymptotic distributions of the event-specific win ratios under proportional hazards (PH) assumptions for the relevant cause-specific hazard functions of the non-terminal and terminal event, respectively. The win ratios converge to the respective hazard ratios under the PH assumptions and therefore are censoring-free, even if the censoring distributions in the two treatment arms may be different. With the asymptotic bivariate normal distributions of the win ratios, confidence intervals and testing procedures can be obtained. With proper transformations the confidence intervals and testing procedures have correct coverage probabilities and type one error rates. The new procedures are illustrated in the clinical trial Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Antagonist (TOPCAT).

Session 84: Statistics in Imaging

Change-point analysis for modern data

♦*Hao Chen*[1]*, Shizhe Chen*[1] *and Xinyi Deng*[2]
[1]University of California, Davis
[2]Beijing University of Technology
hxchen@ucdavis.edu

After observing snapshots of a network, can we tell if there has been a change in dynamics? After collecting spiking activities of thousands of neurons in the brain, how shall we extract meaningful information from the recording? We introduce a change-point analysis framework utilizing graphs representing the similarity among observations. This approach is non-parametric and can be applied to data when an informative similarity measure can be defined. Analytic approximations to the significance of the test statistics are derived to make the method fast applicable to long sequences. The method is illustrated through the analysis of the Neuropixels data.

Quantification in Colocalization Analysis: Beyond "Red+Green=Yellow"

Shulei Wang
University of Illinois at Urbana-Champaign
shuleiw@illinois.edu

Colocalization analysis is a powerful tool to study the associations/interactions between two biomolecules such as proteins via imaging techniques. Despite a diversity of analysis tools developed for colocalization, most of which have notoriously been subject to misinterpretation and inconsistencies due to difficulties in robust quantification and spatial inference. I will share some of our efforts to improve colocalization analysis using statistical and computation tools in this talk.

Methodology for the Comparison of Diffusion Tensor Images Across Cohorts

♦*Gina-Maria Pomann*[1]*, Ana-Maria Staicu*[2] *and Sujit Ghosh*[2]
[1]Duke University
[2]North Carolina State University
gina-maria.pomann@duke.edu

Motivated by a natural history imaging study, we present a nonparametric testing procedure for testing the null hypothesis that two samples of curves observed at discrete grids and with noise have the same underlying distribution. We use functional principal components-based methods to develop a test for the equality of the distributions of two samples of curves, when their eigenfunctions are the same. Our approach reduces the dimensionality of the testing problem in a way that enables the application of traditional nonparametric univariate testing procedures. This results in a procedure that is not only computationally efficient and allows for a variety of sampling designs. This methodology is applied to a diffusion tensor imaging (DTI) study, where the objective is to statistically compare white matter tract profiles between healthy individuals and multiple-sclerosis patients, as assessed by conventional DTI measures.

Longitudinal ComBat: A Method for Harmonizing Longitudinal Multi-scanner Imaging Data

Joanne Beer, Russell Shinohara and ♦*Kristin Linn*
University of Pennsylvania
klinn@pennmedicine.upenn.edu

Neuroimaging is a major underpinning of modern neuroscience research and the study of brain development, abnormality, and disease. Combining neuroimaging datasets from multiple sites and scanners can increase statistical power for detecting biological effects of interest. However, technical variation due to differences in scanner manufacturer, model, and acquisition protocols may bias estimation of these effects. Originally proposed to address batch effects in genomic data set, ComBat has been shown to be effective at removing unwanted variation due to scanner in cross-sectional neuroimaging data. We propose an extension of the ComBat model for longitudinal data and demonstrate its performance using simulations as well as longitudinal cortical thickness data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) stud¡U+0010005C¿. We demonstrate that longitudinal ComBat controls type I error and has higher power for detecting changes in thickness over time compared to alternative methods compared to

naively applying cross-sectional ComBat to the longitudinal thickness trajectories.

Session 85: Advances in false discovery rate control methods

Data splitting methods for FDR controls

Chenguang Dai[1]*, Buyu Lin*[2]*, Xing Xin*[3] *and* ♦*Jun Liu*[4]
[1]Citadel LLC
[2]Harvard University
[3]Virginia Tech
[4]Harvard University
jliu@stat.harvard.edu

A classical statistical idea is to introduce perturbations and examine their impacts on a statistical procedure. We here explore the use of data splitting (DS) for controlling false discovery rate (FDR). A DS procedure simply splits the data into two halves, and computes a statistic reflecting the consistency of the two sets of parameter estimates (e.g., regression coefficients). The FDR control can be achieved by taking advantage of such a statistic. Furthermore, by repeated sample splitting, we propose Multiple Data Splitting (MDS) to stabilize the selection result and boost the power. Interestingly, MDS not only helps overcome the power loss caused by data splitting with the FDR still under control, but also results in a lower variance for the estimated FDR compared with all other methods in consideration. DS and MDS are straightforward conceptually, easy to implement algorithmically, and efficient computationally. Simulation results as well as a real data application show that both DS and MDS control the FDR well and MDS is often the most powerful method among all in consideration, especially when the signals are weak and correlations or partial correlations are high among the features. Our preliminary tests on nonlinear models such as generalized linear models and neural networks also show promises.

Inference with approximate co-sufficient sampling

♦*Rina Barber*[1]*, Lucas Janson*[2] *and Wanrong Zhu*[1]
[1]University of Chicago
[2]Harvard University
rina@uchicago.edu

Goodness-of-fit (GoF) testing is ubiquitous in statistics, with direct ties to model selection, confidence interval construction, conditional independence testing, and multiple testing. While testing the GoF of a simple null hypothesis provides an analyst great flexibility in the choice of test statistic while still ensuring validity, most GoF tests for composite null hypotheses are far more constrained, as the test statistic must have a tractable distribution over the entire null model space. A notable exception is co-sufficient sampling (CSS), which resamples the data conditional on a sufficient statistic to ensure valid Type I error control. But CSS testing requires the null model to have a compact sufficient statistic, which only holds for a very limited class of models; even for a null model as simple as logistic regression, CSS testing is powerless. In this work, we leverage the concept of approximate sufficiency to generalize CSS testing to essentially any parametric model with an asymptotically-efficient estimator; we call our extension "approximate CSS" (aCSS) testing, and establish that it offers high power along with approximate Type I error control in a broad range of settings, including settings where the parameter space is restricted via constraints that enable us to handle sparsity and other problems.

Multiple testing under dependence and non-sparsity

Hongyuan Cao
FSU
hcao@fsu.edu

High throughput technologies enable simultaneous inference of complex high dimensional data. An acute problem is the multiple testing adjustment. Most existing literature examine the problem under independence and sparsity assumptions. We propose a new multiple testing procedure to incorporate dependence and non-sparsity features inherent in many high dimensional data, such as microRNA in genomics and quantitative high throughput screening (qHTS) assays in toxicology. Simulation studies demonstrate the favorable performance of our procedure with competing methods. We illustrate our method with a microRNA dataset.

Interactive identification of individuals with positive treatment effect while controlling false discoveries

♦*Boyan Duan*[1]*, Aaditya Ramdas*[2] *and Larry Wasserman*[2]
[1]Google
[2]Carnegie Mellon University
boyand@google.com

Out of the participants in a randomized experiment with anticipated heterogeneous treatment effects, is it possible to identify which ones have a positive treatment effect, even though each has only taken either treatment or control but not both? While subgroup analysis has received attention, claims about individual participants are more challenging. We frame the problem in terms of multiple hypothesis testing: we think of each individual as a null hypothesis (the potential outcomes are equal, for example) and aim to identify individuals for whom the null is false (the treatment potential outcome stochastically dominates the control, for example). We develop a novel algorithm that identifies such a subset, with nonasymptotic control of the false discovery rate (FDR). Our algorithm allows for interaction - a human data scientist (or a computer program acting on the human's behalf) may adaptively guide the algorithm in a data-dependent manner to gain high identification power. We also propose several extensions: (a) relaxing the null to nonpositive effects, (b) moving from unpaired to paired samples, and (c) subgroup identification. We demonstrate via numerical experiments and theoretical analysis that the proposed method has valid FDR control in finite samples and reasonably high identification power.

Session 86: New classification methods for imaging, network and dynamic treatment

Community Detection in Hypergraphs

Yaoming Zhen and ♦*Junhui Wang*
City University of Hong Kong
j.h.wang@cityu.edu.hk

Network data has attracted tremendous attention in recent years, and most conventional networks focus on pairwise interactions between two vertices. However, real-life network data may display more complex structures, and multi-way interactions among vertices arise naturally, leading to hypergraph networks. In this talk, we will present a novel method for detecting community structure in general hypergraph networks, uniform or non-uniform. It rst introduces a null vertex to augment a non-uniform hypergraph into a uniform multi-hypergraph, and then embeds the multi-hypergraph in a low-dimensional vector space such that vertices within the same community are close to each other. The asymptotic properties of the proposed method will be discussed in terms of both community detection and hypergraph estimation, which are also supported by numerical experiments on some simulated and real examples.

High-order Joint Embedding for Multi-Level Link Prediction

♦*Yubai Yuan and Annie Qu*
University of California, Irvine
yubaiy@uci.edu

Link prediction infers potential links from observed networks, and is one of the essential problems in network analyses. In contrast to traditional graph representation modeling which only predicts two-way pairwise relations, we propose a novel tensor-based joint network embedding approach on simultaneously encoding pairwise links and hyperlinks onto a latent space, which captures the dependency between pairwise and multi-way links in inferring the potential unobserved hyperlinks. The major advantage of the proposed embedding procedure is that it incorporates both the pairwise relationships and subgroup-wise structure among nodes to capture richer network information. In addition, the proposed method introduces a hierarchical dependency among links to allow hyperlink augmentation, and leads to better link prediction. In theory we establish the estimation consistency for the proposed embedding approach, and provide a faster convergence rate compared to link prediction utilizing pairwise links or hyperlinks only. Numerical studies on simulation settings and Facebook ego-networks indicate that the proposed method improves hyperlink and pairwise link prediction accuracy compared to existing link prediction algorithms.

Solving Infinite Horizon Dynamic Treatment Regimes: A pT Learning Framework

♦*Wenzhuo Zhou*[1]*, Ruoqing Zhu*[1] *and Annie Qu*[2]
[1]University of Illinois Urbana Champaign
[2]University of California Irvine
wenzhuo3@illinois.edu

Recent advances in mobile technology provide an effective way to monitor patients' health status and deliver real-time adaptive interventions to patients with chronic disease. In general, the mobile health (mHealth) interventions can be formalized as a dynamic treatment regime (DTR), which consists of a sequential decision rule in maximizing the long-term benefits. Most existing approaches are proposed to estimate the optimal DTRs over a fixed short-time period. However, many mHealth applications are designed for the extremely long-term use and thus required to estimate optimal DTR beyond the time horizon of the collected data. We propose a proximal temporal consistency learning (pT-Learning) framework to obtain a constrained minimax DTR estimator for estimating a sparse and near-optimal DTR. The method utilize a smoothed objective function while can easily incorporate the off-policy data without inverse probability weighting. We show that the proposed estimator can be solved using an efficient and scalable stochastic gradient descent algorithm. We establish the excess risk bound, and a finite sample error bound for the estimated value function. The numerical performance is evaluated through simulation studies, and the OhioT1DM type 1 diabetes dataset.

Dermoscopic Image Classification with Neural Style Transfer

♦*Yutong Li*[1]*, Ruoqing Zhu*[1]*, Mike Yeh*[1] *and Annie Qu*[2]
[1]University of Illinois at Urbana-Champaign
[2]University of California, Irvine
li228@illinois.edu

Skin cancer, the most commonly found human malignancy, is primarily diagnosed visually via dermoscopic analysis, biopsy, and histopathological examination. However, unlike other types of cancer, automated image classification of skin lesions is deemed more challenging due to the irregularity and variability in the lesions' appearances. In this work, we propose an adaptation of the Neural Style Transfer (NST) as a novel image pre-processing step for skin lesion classification problems. We represent each dermoscopic image as a style image and transfer the style of the lesion onto a homogeneous content image. This transfers the main variability of each lesion onto the same localized region, which allows us to integrate the generated images together and extract latent, low-rank style features via tensor decomposition. We train and cross-validate our model on a dermoscopic data set collected and preprocessed from the International Skin Imaging Collaboration (ISIC) database. We show that the classification performance based on the extracted tensor features using the style-transferred images significantly outperforms that of the raw images by more than 10%, and is also competitive with well-studied, pre-trained CNN models using transfer learning. Additionally, the tensor decomposition further identifies latent style clusters, which may provide clinical interpretation and insights.

Session 87: The Jiann-Ping Hsu Invited Session on Biostatistical and Regulatory Sciences

Statistical Data Analysis in Pharmaceutical Industry

Kao-tai Tsai
BMS
Kao-Tai.Tsai@bms.com

Statistical data analysis is an important function of statisticians in industry; however, it has not been emphasized as it should be. In this session, I will describe the evolution of data analyses through the history of Statistics, the analyses have commonly performed in the pharmaceutical industry, and the machine learning technologies which are very useful in data analyses, especially, in genomics research. Examples from clinical trials will be used to illustrate the applications.

Principles of leading with statistical knowledge and a problem-solving approach to innovation

Stephan Ogenstad
Jiann-Ping Hsu College of Public Health, Georgia Southern University
sogenstad@statogen.com

Innovation is a new idea, method, or device. There are different approaches. One is through lengthy trial and error until one finds a solution to a problem. Another is through systematic methods that are efficient and will lead to a great solution. The approaches can use data or no data. We will look at principles to lead the development with statistical knowledge and problem-solving to reach innovations.

Evaluation of SIMEX extrapolation methods in Accelerated failure time models with covariate measurement error

♦*Manyun Liu, Roshni Modi and Lili Yu*
Georgia Southern University
ml16842@georgiasouthern.edu

It is well known that ignoring measurement errors in covariates in the model leads to biased estimates. Various methods were proposed to address this issue. Simulation and extrapolation (SIMEX) method developed by (He et al. 2007) is a popular method due to its flexibility. It consists of two steps, simulation step and extrapolation step. Although it was investigated for applying to different settings, very little research was done on finding the best extrapolation function. The objective of this study is to evaluate which extrapolation function gives best estimation in AFT models for data follows Weibull distribution. We use simulation studies to investigate the performance of three most popular extrapolation functions used for

the SIMEX method. The results show linear extrapolation function is best for data with small measurement error, quadratic extrapolation function is best for the data with medium measurement error and nonlinear extrapolation function is best for the data with large measurement error. Then we applied these three extrapolation functions to a real data set - patients with advanced lung cancer from the North Central Cancer Treatment Group to illustrate the usefulness of the research.

Session 88: Survey Analysis and Design Using Multilevel Regression and Post-stratification

On the Use of Auxiliary Variables in Multilevel Regression and Poststratification

Yajuan Si
University of Michigan, Ann Arbor
yajuan@umich.edu

Multilevel regression and poststratification (MRP) has become a popular approach for selection bias adjustment in subgroup estimation, with widespread applications from social sciences to public health. We examine the statistical properties of MRP in data integration and inferences of probability and nonprobability-based surveys, and the broad extensions in practical applications. Our development is motivated by the Adolescent Brain Cognitive Development (ABCD) Study that has collected children across 21 U.S. geographic locations for national representation but is subject to selection bias, a common problem of nonprobability samples. The success to MRP prominently depends on the quality of available auxiliary information. In this paper, we develop the statistical foundation of MRP on how to incorporate auxiliary variables. We build up a framework under MRP for statistical data integration and inferences. Our simulation studies indicate the statistical validity of MRP with a tradeoff between robustness and efficiency and present the improvement over alternative methods. We apply the approach to evaluate cognition performances of diverse groups of children in the ABCD study and find that the adjustment of auxiliary variables has a substantial effect on the inference results.

Survey Design for Multilevel Regression and Post-Stratification

♦*Yutao Liu*[1]*, Lauren Kennedy*[2]*, Andrew Gelman*[3] *and Qixuan Chen*[3]
[1]Vertex Pharmaceuticals, Inc.
[2]Monash University
[3]Columbia University
yl3050@caa.columbia.edu

Multilevel regression and post-stratification (MRP) has been widely applied in survey analysis, for both probability and non-probability samples. In spite of the rich literature on survey analysis using MRP, literature on survey design for MRP is scarce. Empirical studies using simulation could be considered in the design stage, but it is computationally intensive. Therefore, it is of interest to develop theoretical results. In this article, we consider survey design for MRP. We propose a general procedure to calculate margin of errors (MOE) with close form formula, given the design parameters, and validate the theoretical results using simulation studies. Simulation results indicate that the procedure yield valid MOE, accounting for partial pooling. We demonstrate the use of the procedure in two survey design scenarios, online panels using quota sampling and telephone surveys with fixed total sample sizes.

Multilevel regression and post-stratification in R

♦*Lauren Kennedy*[1]*, Jonah Gabry*[2]*, Rohan Alexander*[3]*, Dewi Amaliah*[1] *and Mitzi Morris*[2]
[1]Monash University
[2]Columbia University
[3]University of Toronto
lauren.kennedy1@monash.edu

Multilevel regression and post-stratification (MRP) has grown increasingly popular. Rather uniquely, advances to the methodology have matched pace with the spread of applications, leading to simultaneous growth in both use and understanding/modelling developments. We draw upon a framing of MRP to unify the methodological advances into one framework, emphasizing the practical components of an MRP analysis. We then use this framing to develop a package for the statistical software R that implements common versions of an MRP analysis and is easily extensible for new research developments. We hope that this package will aid users in implementing MRP, researchers in distributing their method with less effort than creating an entire new package for each adaption, and continued growth of the field.

Session 89: Massive Data Analysis

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

Xiaoqing Tan, Chung-Chou Chang and ♦*Lu Tang*
University of Pittsburgh
lutang@pitt.edu

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

Integrative Causal Inference in the Presence of Study Heterogeneity

Ling Zhou[1]*,* ♦*Xu Shi*[2]*, Mengtong Hu*[2] *and Peter Song*[2]
[1]Southwestern University of Finance and Econonmics
[2]University of Michigan
shixu@umich.edu

We consider a causal inference problem of practical importance in which multiple similar clinical studies conducted at different sites are merged to derive a meta-estimation for the effect of a treatment of interest with no need to share the raw data from individual studies. This integrative analysis is advantageous to protect data privacy at a high order and to avoid administrative complexities in sharing subject-level information across different study sites. We investigate the bias and variance of the classical inverse-variance weighted average estimator when applied to estimate a causal parameter. We show that when the bias of the meta-estimator is ignorable, the meta-estimator is asymptotically equally efficient compared to the centralized oracle estimator obtained by combining data from all

sites if the study-site data are fully homogeneous, and achieves minimum variance, i.e., more efficient than the centralized oracle estimator, when the study-site data are heterogeneous. When the bias is not ignorable, we propose alternative methods for bias correction. Our findings and proposed methods are illustrated by extensive simulation studies and real-world data examples.

Multivariate Online Regression Analysis with Heterogeneous Streaming Data

♦*Lan Luo*[1] *and Peter Song*[2]

[1]University of Iowa

[2]University of Michigan

lan-luo@uiowa.edu

New data collection and storage technologies have given rise to a new field of streaming data analytics, including real-time statistical methodology for online data analyses. Most existing online learning methods are based on homogeneity assumption such that the sequence of samples are independent and identical. However, inter-data batch correlation and dynamically evolved batch-specific effects are among the key defining features in real-world streaming data such as electronic health records and mobile health data. This talk centers around the state space-mixed models in which the observed data stream is driven by a latent state process that follows a Markov process. In this setting, online maximum likelihood estimation is challenged by high-dimensional integrals and complex covariance structures. In this project, we develop a Kalman filter based real-time regression analysis method that enables to update both point estimates and standard errors of the fixed population average effects while adjusting for dynamic hidden effects. Both theoretical justification and numerical experiments have demonstrated that our proposed online method has similar statistical properties to its offline counterpart but enjoys great computation efficiency. We also apply this method to analyze an electronic health record data example.

Center-augmented l2-type regularization for subgroup learning

♦*Ling Zhou*[1]*, Ye He*[2]*, Huazhen Lin*[1] *and Yingcun Xia*[3]

[1]Southwestern University of Finance and Economics

[2]University of Electronic Science and Technology of China

[3]National University of Singapore

zhouling@swufe.edu.cn

The existing methods for subgroup analysis can be roughly divided into two categories: finite mixture models (FMM) and regularization methods with an l1- type penalty. In this paper, by introducing the group centers and l2-type penalty in the loss function, we propose a novel centre-augmented regularization (CAR) method; this method can be regarded as a unification of the regularization method and FMM and hence exhibits higher efficiency and robustness and simpler computations than the existing methods. Particularly, its computational complexity is reduced from the O(n2) of the conventional pairwise-penalty method to only O(nK), where n is the sample size and K is the number of subgroups. The asymptotic normality of CAR is established, and the convergence of the algorithm is proven. CAR is applied to a dataset from a multi-center clinical trial: Buprenorphine in the Treatment of Opiate Dependence; a larger R2 is produced and three additional significant variables are identified compared to those of the existing methods.

Session 90: Modern streaming data analysis: change-point problems and applications

Detection of Multiple Transient Changes

♦*Michael Baron*[1] *and Sergey Malov*[2]

[1]American University

[2]St. Petersburg State University,

baron@american.edu

Change-point detection methods are proposed for the case of transient changes, when an unexpected disorder is ultimately followed by an adjustment and return to the initial state. A known base distribution of the in-control state changes to an out-of-control distribution only temporarily. Durations of out-of-control segments vary, but the eventual return to the base distribution is inevitable. Under this general framework, we propose change detection algorithms that control the familywise false alarm and false adjustment rates and derive results on the accuracy of the obtained change-point estimates. Examples of similar problems are shown in quality and process control, energy finance, and statistical genetics, although the meaning of disorder and adjustment change-points is quite different in these applications. [Acknowledgment: The work of M. Baron is supported by U.S. National Science Foundation grant 1737960. The work of S. Malov is supported by the Russian Science Foundation grant 20-14-00072.]

Equivariant Variance Estimation for Multiple Change-point Model

♦*Ning Hao*[1]*, Yue Niu*[1] *and Han Xiao*[2]

[1]University of Arizona

[2]Rutgers University

nhao@math.arizona.edu

The variance of noise plays an important role in many change-point detection procedures and the associated inferences. Most commonly used variance estimators require strong assumptions on the true mean structure or normality of the error distribution, which may not hold in applications. In this talk, we introduce a framework of equivariant variance estimation for multiple change-point models. In particular, we characterize the set of all equivariant unbiased quadratic variance estimators for a family of change-point model classes, and develop a minimax theory for such estimators.

Changepoint detection in autocorrelated ordinal categorical time series

Mo Li and ♦*QiQi Lu*

Virginia Commonwealth University

qlu2@vcu.edu

While changepoint aspects in correlated sequences of continuous random variables have been extensively explored in the literature, changepoint methods for independent categorical time series are only now coming into vogue. This talk extends changepoint methods by developing techniques for serially correlated categorical time series. In this study, a cumulative sum type test is devised to test for a single changepoint in a correlated categorical data sequence. Our categorical series is constructed from a latent Gaussian process through clipping techniques. A sequential parameter estimation method is proposed to estimate the parameters in this model. The methods are illustrated via simulations and applied to a real categorized rainfall time series from Albuquerque, New Mexico.

Univariate Likelihood Projections and Characterizations of the Multivariate Normal Distribution Applicable to a Multivariate Change Point Detection

Albert Vexler

Professor
avexler@buffalo.edu

The problem of characterizing a multivariate distribution of a random vector using examination of univariate combinations of vector components is an essential issue of multivariate analysis. The likelihood principle plays a prominent role in developing powerful statistical inference tools. In this context, we raise the question: can the univariate likelihood function based on a random vector be used to provide the uniqueness in reconstructing the vector distribution? In multivariate normal (MN) frameworks, this question links to a reverse of Cochran's theorem that concerns the distribution of quadratic forms in normal variables. We characterize the MN distribution through the univariate likelihood type projections. The proposed principle is employed to illustrate simple techniques for testing the hypothesis: "observed vectors are from a MN distribution" versus that "firs data points are from a MN distribution, and then, starting from an unknown position, observations are non-MN distributed". In this context, the proposed characterizations of MN distributions allow us to employ well-known mechanisms that use univariate observations. The displayed testing strategy can exhibit high and stable power characteristics, when observed vectors satisfy the alternative hypothesis, whereas their components are normally distributed random variables. In such cases, classical change point detections based on, e.g., Shapiro-Wilk, Henze-Zirklers and Mardia type engines, may break down completely.

Session 91: Recent advances in statistical methods for large-scale omics data

A likelihood-based approach for multivariate categorical response regression in high dimensions

♦*Aaron Molstad*[1] *and Adam Rothman*[2]
[1]University of Florida
[2]University of Minnesota
amolstad@ufl.edu

We propose a penalized likelihood method to fit the bivariate categorical response regression model. Our method allows practitioners to estimate which predictors are irrelevant, which predictors only affect the marginal distributions of the bivariate response, and which predictors affect both the marginal distributions and log odds ratios. To compute our estimator, we propose an efficient algorithm which we extend to settings where some subjects have only one response variable measured, i.e., a semi-supervised setting. We derive an asymptotic error bound which illustrates the performance of our estimator in high-dimensional settings. Generalizations to the multivariate categorical response regression model are proposed. Finally, simulation studies and an application in pan-cancer risk prediction demonstrate the usefulness of our method in terms of interpretability and prediction accuracy.

Estimating gene-to-trait effect with GWAS and eQTL summary statistics using Bayesian Hierarchical Models

♦*ANQI ZHU*[1]*, Nana Matoba*[2]*, Emma Wilson*[3]*, Amanda Tapia*[4]*, Yun Li*[5]*, Joseph Ibrahim*[4]*, Jason Stein*[2] *and Michael Love*[6]
[1]Department of Biostatistics University of North Carolina at Chapel Hill, 23andMe Inc.
[2]Neuroscience Center, Department of Genetics, University of North Carolina at Chapel Hill
[3]Department of Genetics, University of North Carolina at Chapel Hill
[4]Department of Biostatistics, University of North Carolina at Chapel Hill
[5]Department of Biostatistics, Department of Genetics and Department of Computer Science, University of North Carolina at Chapel Hill
[6]Department of Biostatistics and Department of Genetics, University of North Carolina at Chapel Hill
anqiz@23andme.com

Expression quantitative trait loci (eQTL) studies are used to understand the regulatory function of non-coding genome-wide association study (GWAS) risk loci, if the gene has a role as a mediator of the complex trait or disease. Loci that exhibit allelic heterogeneity, that is, loci containing multiple causal variants, offer the opportunity to investigate whether effects are concordant and proportional across eQTL and GWAS; if the gene is a partial mediator of the trait, the sign and size of the effects across distinct eQTL variants should be reflected in GWAS associations. Here, we introduce a new statistical method, MRLocus, for Bayesian estimation of the gene-to-trait effect from eQTL and GWAS summary data for loci with evidence of allelic heterogeneity. MRLocus makes use of a colocalization step applied to each nearly-LD-independent eQTL, followed by an MR analysis step across eQTLs. Additionally, our method involves estimation of the extent of allelic heterogeneity through a dispersion parameter, indicating variable mediation effects from each individual eQTL on the downstream trait. In simulation, the method has higher accuracy and better uncertainty measures compared to other competing methods, and we compare its estimates on candidate causal gene-trait pairs from literature.

Developing Trans-ethnic Polygenic Risk Scores Using Empirical Bayes and Super Learning Algorithm

Haoyu Zhang
Harvard T.H. Chan School of Public Health
haoyuzhang@hsph.harvard.edu

Polygenic risk scores (PRS) are useful for predicting various phenotypes/outcomes; however, as most PRS have been developed with data generated in European Ancestry (EA) populations, performances of PRS are often poorer in non-EA populations, reflecting their degree of divergence from EA population. To improve PRS performance in non-EA populations, we propose a novel method, Two-Dimensional Clumping and Thresholding with Super Learning and Empirical Bayes (TDLD-SLEB), which takes advantage of both existing large GWAS from EA populations and smaller GWAS from non-EA populations. TDLD-SLEB uses a two-dimensional thresholding method to incorporate SNPs that have either effects in both the larger (e.g., EA populations) and the smaller (e.g., non-EA populations) target population or specific effects in the smaller population. It estimates effect sizes for SNPs in the target population using an Empirical-Bayes method that borrows GWAS information from across all populations. Finally, it incorporates a super learning algorithm to combine the series of PRS generated by the various SNP selection thresholds for the target population. Our simulation analyses mimicked real LD patterns using haplotype data of 1000 Genomes Project (Phase 3) for five ancestries. We considered various genetic architectures including different levels of negative selection and genetic correlation across ancestries. We found PRSs generated by TDLD-SLEB had significantly improved predictionary accuracy for non-EA populations in independent validation datasets, compared to single ethnic PRS, EUR derived PRS, or a weighted PRS that combines EUR and single ethnic derived PRS with weights selected to optimize prediction in the target population. Using data from 23andMe, Inc., we developed and vali-

dated population specific PRS for seven complex traits using GWAS data from Europeans (average N=2,442K), African American (average N=113K), Latino (average N=411K), East Asians (average N=94K), South Asians (average N=25K). We found TDLD-SLEB often led to large improvement in performance of PRS compared to alternative methods for predicting traits in the African American population (e.g., average R^2 increased +277% compared to the weighted PRS method). For other ethnic groups, TDLD-SLEB also led to sometime notable improvements in the performance of PRS, such as for the cardiovascular disease in the Latino population (AUC = 0.61 for TDLD-SLEB vs. AUC= 0.58 for the weighed PRS method). In conclusion, we developed a computationally scalable and statistically efficient method for generating predictive PRS in non-European populations using GWAS datasets across diverse populations.

De-biased Lasso for Generalized Linear Models with A Diverging Number of Covariates

♦*Lu Xia*[1]*, Bin Nan*[2] *and Yi Li*[3]
[1]University of Washington
[2]University of California, Irvine
[3]University of Michigan
xialu@uw.edu

Modeling and drawing inference on the joint associations between single nucleotide polymorphisms and a disease has sparked interest in genome-wide associations studies. In the motivating Boston Lung Cancer Survival Cohort (BLCSC) data, a large number of single nucleotide polymorphisms of interest, though still smaller than the sample size, challenges inference on their joint associations with the disease outcome. In similar settings, we find that neither the debiased lasso approach (van de Geer et al., 2014), which assumes sparsity on the inverse information matrix, nor the standard maximum likelihood method can yield confidence intervals with satisfactory coverage probabilities for generalized linear models. Under this "large n, diverging p" scenario, we propose an alternative debiased lasso approach by directly inverting the Hessian matrix without imposing the matrix sparsity assumption, which further reduces bias compared to the original de-biased lasso and ensures valid confidence intervals with nominal coverage probabilities. We establish the asymptotic distributions of any linear combinations of the parameter estimates, which lays the theoretical ground for drawing inference. Simulations show that the proposed refined de-biased estimating method performs well in removing bias and yields honest confidence interval coverage. We use the proposed method to analyze the aforementioned BLCSC data, a large scale hospital-based epidemiology cohort study, that investigates the joint effects of genetic variants on lung cancer risks.

Session 92: Mixtures, two-sample problems and their applications

Testing Homogeneity in Contaminated Mixture Models

Guanfu Liu[1]*,* ♦*Yuejiao (Cindy) Fu*[2]*, Wenchen Liu*[3] *and Rongji Mu*[4]
[1]Shanghai University of International Business and Economics
[2]York University
[3]Shanghai Lixin University of Accounting and Finance
[4]Shanghai Jiao Tong University
yuejiao@yorku.ca

Contaminated mixture models (CMMs) have wide applications in the real world. Testing homogeneity in the CMMs is an interesting and important research problem. In this paper, we develop an EM-test for homogeneity in the general framework of the CMMs. The null limiting distribution of the test is shown to be a shifted mixture of chi-square distributions. Simulation studies demonstrate that the EM-test has excellent finite-sample performance. Two real-data examples illustrate the applications of the proposed method.

Statistical inference for a relaxation index of stochastic dominance under density ratio model

Weiwei Zhuang
University of Science and Technology of China
weizh@ustc.edu.cn

Stochastic dominance is usually used to rank random variables by comparing their distributions, so it is widely applied in economics and finance. In actual applications, complete stochastic dominance is too demanding to meet, so relaxation indexes of stochastic dominance have attracted more attention. The p index, the biggest gap between two distributions, can be a measure of the degree of deviation from complete dominance. The traditional estimation method is to use the empirical distribution functions to estimate it. Considering the populations under comparison are generally of the same nature, we can link the populations through density ratio model under certain condition. Based this model, we propose a new estimator and establish its statistical inference theory. Simulation results show that the proposed estimator substantially improves estimation efficiency and power of the tests, and coverage probabilities satisfactorily match the confidence levels of the tests, which show the superiority of the proposed estimator. Finally we apply our method to a real example of the Chinese household incomes.

Estimation of B-spline copula

Xiaoling Dou
Waseda University
dou@waseda.jp

The B-spline copula is a generalization of the Bernstein copula. It is defined by replacing the Bernstein basis functions by B-spline basis functions. This change requires the copula parameters satisfy slightly different conditions, in spite of the copula form remains the same. Because the Bernstein copula can be considered as a finite mixture distribution for given marginals, we can use EM algorithm methods to estimate the Bernstein copula. Since this idea is also available for the B-spline copula, we propose to generate the existing EM algorithms of the Bernstein copula to estimate the B-spline copula by changing the basis functions and the parameter conditions. When data are highly dependent, we can also consider the copula is a mixture of two special copulas, the B-spline copula with maximum correlation and the independent B-spline copula. In this case, the model has less parameters and a grid method for estimating the parameters is easy to use. Illustrative examples are presented with real data sets.

Stochastic Simulation Models for the Spread of Infectious Diseases

Zhiyong Jin, Lin Xue and ♦*Liqun Wang*
University of Manitoba
Liqun.Wang@umanitoba.ca

The widely used epidemic models, such SEIR and its various generalizations, are defined through a set of differential equations and the transition rates between compartments are often assumed to be constant. However, in practice more realistic infectivity profiles are often time-varying, partly because of public health measures and knowledge of new diseases. For example, the total contact rates among a population are time-varying as the social patterns of con-

tact in a particular society change. In addition, they can also be different among different populations or ethnic communities. In this paper, we propose a stochastic simulation model that provides a very flexible framework compared with the traditional deterministic compartmental models in epidemiology. For example, we simulate various social patterns of contacts, such as uniform and independent movement, independent Brownian Motion, and Levy flight pattern. We also apply a social network to model the spread of a disease and the potential loss incurred by an epidemic. Similar to our proposed simulation model, the network model also provides a more flexible framework compared with the deterministic compartmental models. Since a complete network model requires the information of all individuals and their links to be known in advance, how to properly design a network that is close to reality is of primary interest. We show a simulation example using the network epidemic model.

Session 93: Statistical methods for complex data

Functional Data Analysis Using Neural Networks

Sneha Jadhav
Wake Forest
jadhavs@wfu.edu

Functional Data Analysis typically involves data observed over time or space. Though several methods have been developed to investigate several aspects of functional data, despite being well suited for the problem neural networks are not commonly used. We explore the use of different neural network architectures in functional data analyses.

Bias-corrected estimation in errors-in-variables Logistic regression under case-control study

♦*Pei Geng and Huyen Nguyen*
Illinois State University
pgeng@ilstu.edu

It was discovered by Qin and Zhang (1997) that the logarithm ratio of the conditional covariate density for the case and control groups in the Logistic regression is a linear function. Geng and Sakhanenko (2016) developed parameter estimation in Logistic regression by minimizing the integrated square distance (ISD) based on estimated density log-ratio. However, when the covariate is collected with measurement errors, the naïve estimator based on ISD suffers from asymptotic bias. A bias-corrected estimation approach is proposed by adapting the deconvolutional kernel density estimators. The consistency and asymptotic normality of the proposed estimator are derived. Simulation study shows superior performance of the bias-corrected estimation. The proposed method is also applied to the Framingham Heart Study.

A unified framework for change point detection in high-dimensional linear models

♦*Abolfazl Safikhani and Yue Bai*
University of Florida
a.safikhani@ufl.edu

Detecting structural breaks in high-dimensional linear regression models is a challenging task due to the existence of an unknown number of such breaks, unknown location of breaks, and unknown high-dimensional model parameters. A general methodology for handling this problem will be presented which can cover a wide range of statistical models including mean shift models, Vector Auto-Regressive Models (VARs), and Gaussian graphical models. The proposed algorithm consists of (1) applying blocked fused lasso to identify potential locations of breaks; (2) evaluate the magnitude of changes and apply thresholding to keep only the large enough jumps; (3) apply k-means clustering to the location of large jumps to select clusters around true breaks; (4) exhaustive search within each selected cluster to locate the breaks. Consistency for estimating the number of breaks, their locations and model parameters is verified under mild conditions satisficed in many well-known high-dimensional statistical models. The proposed method performs well in synthetic and real data applications while outperforming some competing methods in the literature.

Rapid Online Plant Leaf Area Change Detection with High-Throughput Plant Image Data

Yinglun Zhan, Ruizhi Zhang, ♦*Yuzhen Zhou, Vincent Stoerger, Jeremy Hiller, Tala Awada and Yufeng Ge*
University of Nebraska Lincoln
yuzhenzhou@unl.edu

High-throughput plant phenotyping (HTPP) has become an emerging technique to study plant traits due to its fast, labor saving, accurate, and non-destructive nature. It has wide applications in plant breeding and crop management. However, the resulting massive image data has raised up a challenge associated with efficient plant traits prediction and anomaly detection. In this paper, we propose a two-step image-based online detection framework for monitoring and quick change detection of the individual plant leaf area via real-time imaging data. Our proposed method is able to achieve smaller detection delay compared with some baseline methods under some predefined false alarm rate constraint. Moreover, it does not need to store all past image information and can be implemented in real-time. The efficiency of the proposed framework is validated by a real data analysis.

Session 94: Frontiers in Semi-Parametric and Non-Parametric Methods

Statistical inference for functional time series: autocovariance function

Chen Zhong and ♦*Lijian Yang*
Tsinghua University
yanglijian@mail.tsinghua.edu.cn

Statistical inference for functional time series is investigated by extending the classic concept of autocovariance function (ACF) to functional ACF (FACF). It is established that for functional moving average (FMA) data, the FMA order can be determined as the highest nonvanishing order of FACF, just as in classic time series analysis. A two-step estimator is proposed for FACF, the first step involving simultaneous B-spline estimation of each time trajectory and the second step plug-in estimation of FACF by using the estimated trajectories in place of the latent true curves. Under simple and mild assumptions, the proposed tensor product spline FACF estimator is asymptotically equivalent to the oracle estimator with all known trajectories, leading to asymptotic correct simultaneous confifidence envelope (SCE) for the true FACF. Simulation experiments validate the asymptotic correctness of the SCE and data-driven FMA order selection. The proposed SCEs are computed for the FACFs of an ElectroEncephalogram (EEG) functional time series with interesting discovery of fifinite FMA lag and Fourier form functional principal components.

Estimation of the Mean Function of Functional Data via Deep Neural Networks

Shuoyang Wang[1], ♦*Guanqun Cao*[1] *and Zuofeng Shang*[2]
[1]Auburn University

[2]New Jersey Institute of Technology
gzc0009@auburn.edu

In this work, we propose a deep neural networks based method to perform nonparametric regression for functional data. The proposed estimators are based on sparsely connected deep neural networks with ReLU activation function. We provide the convergence rate of the proposed deep neural networks estimator in terms of the empirical norm. We discuss how to properly select of the architecture parameters by cross-validation. Through Monte Carlo simulation studies we examine the finite-sample performance of the proposed method. Finally, the proposed method is applied to analyze positron emission tomography images of patients with Alzheimer disease obtained from the Alzheimer Disease Neuroimaging Initiative database.

Two-Step Time Series Modelling

Lijian Yang[1], ♦*Qin Shao*[2], *Yuanyuan Zhang*[1], *Hanh Nguyen*[2], *Rawiyah Alraddadi*[2] *and Rong Liu*[2]
[1]Tsinghua University
[2]University of Toledo
Qin.Shao@UToledo.Edu

Time series often contain a trend, and detrending is the first step to conduct statistical analysis for such observations. Statistical inference is usually implemented to detrended series. This two-step approach has a long history and is applied to a broad range of time series data. Our work aims at providing theoretical justification for statistical inference based on detrended time series when the trend is estimated nonparametrically from certain types of mathematical base functions. The focus is to illustrate its computational simplicity and theoretical soundness of the two-step approach by several examples.

Free-knot Splines for Generalized Regression Models

♦*Jing Wang and Elena Graetz*
University of Illinois at Chicago
jiwang12@uic.edu

It is well known that Free-knot spline introduces flexible location parameters in the polynomial splines in order to capture the nonlinear trend and distinct structure changes. The free location unknowns bring advantages to capture marked data pattern precisely and also unavoidable lethargy and heavy computational cost. In this paper, we proposed a free-knot spline confidence bands via penalized likelihood function maximization procedures. To optimize the complex objective functions efficiently, we borrow the Automatic Differentiation Model Builders proposed by Skaug and Fourier (2006) to approximate the general likelihood functions. Wild bootstrapping algorithms is adopted in order to obtain consistent variance estimates and construct confidence bands in a generalized regression models when variance is a function of its mean. With satisfied performance and improved computation power, the proposed methods have been applied in a general model setting ups in simulation studies and a couple of real data analysis.

Session 95: Statistical Challenges for Single-cell RNA Sequencing Data Analysis

On the strategy for supervised cell type identification in single-cell RNA-seq

Wenjing Ma, Kenong Su and ♦*Hao Wu*
Emory University
hao.wu@emory.edu

The cell type identification is one of the most important questions in single-cell RNA sequencing (scRNA-seq) data analysis. With the accumulation of public scRNA-seq data, the supervised cell type identification methods have gained increasing popularity due to the better accuracy, robustness, and computational performance. Despite all the advantages, the performance of the supervised methods relies heavily on several key factors: feature selection, prediction method, and most importantly, choice of the reference dataset. In this work, we perform extensive real data analyses to systematically evaluate these strategies in supervised cell identification. We first benchmark nine classifiers along with six feature selection strategies and investigate the impact of reference data size and number of cell types in cell type prediction. Next, we focus on how discrepancies between reference and target datasets would affect the prediction performance. We also investigate the strategies of pooling and purifying reference data. Based on our analysis results, we provide a guideline and rule of thumb for using the supervised cell typing methods: we suggest combining all individuals from available datasets to construct the reference dataset and use Multi-Layer Perceptron (MLP) as the classifier along with F-test as the feature selection method. All the code used for our analysis is available on GitHub (https://github.com/marvinquiet/Celltyping$_{RefConstruct}$).

Semi-supervised learning for Single Cell Multi-Omics Data

Wei Chen
University of Pittsburgh
wec47@pitt.edu

Droplet-based single cell transcriptome sequencing (scRNA-seq) technology is able to measure gene expression from tens of thousands of single cells simultaneously. More recently, coupled with the cutting-edge Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq) and cell hashing, the droplet-based system has allowed for immunophenotyping of single cells based on cell surface expression (ADT data) of specific proteins together with simultaneous transcriptome profiling in the same cell. Different from scRNA-seq data, surface protein (ADT) data don't have limitations such as drop-out issue and thus are capable of labeling many common cell types such as B cells, T cells and NK cells through popular tools like gating. In this study, we developed a novel ADT-guided semi-supervised clustering and classification approach for joint analyzing CITE-seq data and scRNA-seq data from the same or similar tissue types. The novel features of our method include: 1) robustness to cell label noises from ADT data, 2) improving cell clustering performance with the additional scRNA-seq data and the common cell type guidance from ADT data, 3) predicting common cell type labels for scRNA-seq data without ADT data and 4) modeling the consistency/inconsistency between ADT data and RNA data. We demonstrate the validity and usefulness of our method through extensive simulation studies and real data applications. We believe our method will be a useful tool in single cell research community and facilitate novel biological discoveries.

A graph neural network model to estimate cell-wise metabolic flux using single cell RNA-seq data

Wennan Chang[1], *Norah Alghamdi*[2], *Sha Cao*[2] *and* ♦*Chi Zhang*[2]
[1]Purdue University
[2]Indiana University
czhang87@iu.edu

The metabolic heterogeneity, and metabolic interplay between cells and their microenvironment have been known as significant contributors to disease treatment resistance. However, with the lack

of a mature high-throughput single cell metabolomics technology, we are yet to establish systematic understanding of intra-tissue metabolic heterogeneity and cooperation phenomena among cell populations. To mitigate this knowledge gap, we developed a novel computational method, namely scFEA (single cell Flux Estimation Analysis), to infer single cell fluxome from single cell RNA-sequencing (scRNA-seq) data. scFEA is empowered by a comprehensively reconstructed human metabolic map into a factor graph, a novel probabilistic model to leverage the flux balance constraints on scRNA-seq data, and a novel graph neural network based optimization solver. The intricate information cascade from transcriptome to metabolome was captured using multi-layer neural networks to fully capitulate the non-linear dependency between enzymatic gene expressions and reaction rates. We experimentally validated scFEA by generating an scRNA-seq dataset with matched metabolomics data on cells of perturbed oxygen and genetic conditions. Application of scFEA on this dataset demonstrated the consistency between predicted flux and metabolic imbalance with the observed variation of metabolite abundance in the matched metabolomics data. We also applied scFEA on five publicly available scRNA-seq and spatial transcriptomics datasets and identified context and cell group specific metabolic variations. The cell-wise fluxome predicted by scFEA empowers a series of downstream analysis including identification of metabolic modules or cell groups that share common metabolic variations, sensitivity evaluation of enzymes with regards to their impact on the whole metabolic flux, and inference of cell-tissue and cell-cell metabolic communications.

Scaffold: a data generation simulation framework for single-cell RNA-seq data

Rhonda Bacher
University of Florida
rbacher@ufl.edu

Single cell RNA-sequencing (scRNA-seq) is a promising tool that facilitates study of the transcriptome at the resolution of a single cell. However, along with the many advantages of scRNA-seq come technical artifacts not observed in bulk RNA-seq studies including an abundance of zeros, varying levels of technical bias across gene groups, and systematic variation in the relationship between sequencing depth and gene expression (count-depth relationship). To investigate the source of this variability, we developed, scaffold, a first principles simulation framework which takes each step of generating scRNA-seq data into account. With this framework, we demonstrate the contribution of various protocol choices to technical artifacts observed in scRNA-seq data. Furthermore, we illustrate how a critical step in scRNA-seq protocols directly contributes to the systematic variability in the count depth relationship, and show that hypotheses generated with the simulation are supported by existing independent datasets.

Session 96: Semiparametric statistical inference and application

The profile likelihood based statistical inference

Ziqi Chen
East China Normal University
zqchen@fem.ecnu.edu.cn

Motivated by an entropy inequality, we propose the new profile likelihood framework for statistical inference for the multiple linear regression models and the longitudinal data models. The simulation studies and asymptotic properties confirm the superior performances of the proposed methods.

Classified Generalized Linear Mixed Model Prediction Incorporating Pseudo-prior Information

♦*Haiqiang Ma*[1] *and Jiming Jiang*[2]
[1]Jiangxi University of Finance and Economics
[2]University of California, Davis
mhqtc118@163.com

We develop a method of classified mixed model prediction based on generalized linear mixed models that incorporates pseudo-prior information to improve prediction accuracy. We establish consistency of the proposed method both in terms of prediction of the true mixed effect of interest and in terms of correctly identifying the potential class corresponding to the new observations, if such a class exists that matches one of the training data classes. Empirical results, including simulation studies and real-data validation, fully support the theoretical findings.

Nonparametric Regression with Covariates Subject to Dependent Censoring

Hui Jiang[1], ♦*Lei Huang*[2] *and Yingcun Xia*[3]
[1]Huazhong University of Science and Technology
[2]Southwest Jiaotong University
[3]National University of Singapore
stahl@swjtu.edu.cn

Censoring occurs often in data collection. In this paper, we consider nonparametric regression when the covariate is censored under general settings. In contrast to censoring in the response variable in survival analysis, regression with censored covariates is more challenging. We propose to estimate the regression function using conditional hazard rates. Asymptotic normality of our proposed estimator is established. Both theoretical results and simulation studies demonstrate that the proposed method is more efficient than that based on complete observations and other methods, especially when the censoring rate is high. We llustrate the usefulness of the method using a well-known dataset from a randomized placebo controlled clinical trial of the drug D-penicillamine.

Session 97: Advances in High-dimensional Statistics

Variable Selection for Frechet Regression

Yichao Wu
University of Illinois at Chicago
yichaowu@uic.edu

In this talk, I will talk about new variable selection methods for both global Frechet regression and local Frechet regression.

Understanding Generalization in Deep Learning via Tensor Methods

Jingling Li[1], *Yanchao Sun*[1], *Jiahao Su*[1], *Taiji Suzuki*[2] *and* ♦*Furong Huang*[1]
[1]UMD
[2]Riken, Japan
furongh@umd.edu

Deep neural networks generalize well on unseen data though the number of parameters often far exceeds the number of training examples. Recently proposed complexity measures have provided insights to understanding the generalizability in neural networks from perspectives of PAC-Bayes, robustness, overparametrization, compression and so on. In this work, we advance the understanding of the relations between the network's architecture and its generalizability from the compression perspective. Using tensor anal-

ysis, we propose a series of intuitive, data-dependent and easily-measurable properties that tightly characterize the compressibility and generalizability of neural networks; thus, in practice, our generalization bound outperforms the previous compression-based ones, especially for neural networks using tensors as their weight kernels (e.g. CNNs). Moreover, these intuitive measurements provide further insights into designing neural network architectures with properties favorable for better/guaranteed generalizability. Our experimental results demonstrate that through the proposed measurable properties, our generalization error bound matches the trend of the test error well. Our theoretical analysis further provides justifications for the empirical success and limitations of some widely-used tensor-based compression approaches. We also discover the improvements to the compressibility and robustness of current neural networks when incorporating tensor operations via our proposed layer-wise structure.

Bidimensional Linked Matrix Decomposition for Pan-Omics Pan-Cancer Analysis

ERIC LOCK
UNIVERSITY OF MINNESOTA
elock@umn.edu

Several recent methods address the integrative dimension reduction and decomposition of linked high-content data matrices. Typically, these methods consider one dimension, rows or columns, that is shared among the matrices. This shared dimension may represent common features measured for different sample sets (horizontal integration) or a common sample set with features from different platforms (vertical integration). This is limiting for data that take the form of bidimensionally linked matrices, e.g., multiple molecular omics platforms measured for multiple sample cohorts, which are increasingly common in biomedical studies. We propose a flexible approach to the simultaneous factorization and decomposition of variation across bidimensionally linked matrices, BIDIFAC+. This decomposes variation into a series of low-rank components that may be shared across any number of row sets (e.g., omics platforms) or column sets (e.g., sample cohorts). Our objective function extends nuclear norm penalization, is motivated by random matrix theory, and can be shown to give the mode of a Bayesian posterior distribution. We apply the method to pan-omics pan-cancer data from The Cancer Genome Atlas (TCGA), integrating data from 4 different omics platforms and 29 different cancer types.

Tensor Factor Model Estimation by Iterative Projection

Yuefeng Han[1]*, Rong Chen*[1]*,* ♦*Dan Yang*[2] *and Cun-Hui Zhang*[1]
[1]Rutgers University
[2]The University of Hong Kong
dyanghku@hku.hk

Tensor time series, which is a time series consisting of tensorial observations, has become ubiquitous. It typically exhibits high dimensionality. One approach for dimension reduction is to use a factor model structure, in a form like Tucker tensor decomposition, except that the time dimension is treated as a dynamic process with a time dependent structure. In this talk we introduce two approaches to estimate such a tensor factor model by using iterative orthogonal projections of the original tensor time series. These approaches extend the existing estimation procedures and improve the estimation accuracy and convergence rate significantly as proven in our theoretical investigation. Our algorithms are similar to the higher order orthogonal projection method for tensor decomposition, but with significant differences due to the need to unfold tensors in the iterations and the use of autocorrelation. Computational and statistical lower bounds are derived to prove the optimality of the sample size requirement and convergence rate for the proposed methods. Simulation study is conducted to further illustrate the statistical properties of these estimators.

Session 98: Factor Analysis and Random Matrix Theory

Eigen selection in spectral clustering: a theory guided practice

Xiao Han[1]*, Xin Tong*[2] *and* ♦*Yingying Fan*[2]
[1]USTC
[2]USC
fanyingy@usc.edu

Based on a Gaussian mixture type model of K components, we derive eigen selection proce- dures that improve the usual spectral clustering algorithms in high-dimensional settings, which typically act on the top few eigenvectors of an affinity matrix (e.g., $X^T X$) derived from the data matrix X. Our selection principle formalizes two intuitions: (1) eigenvectors should be dropped when they have no clustering power; (2) some eigenvectors corresponding to smaller spiked eigenvalues should be dropped due to estimation inaccuracy. Our selection procedures lead to new spectral clustering algorithms: ESSC for K = 2 and GESSC for K>2. The newly proposed algorithms enjoy better stability and compare favorably against canonical alternatives, as demonstrated in extensive simulation and multiple real data studies.

Selecting the number of components in PCA via random signflips

David Hong, Yue Sheng and ♦*Edgar Dobriban*
UPenn
dobribanedgar@gmail.com

Dimensionality reduction via PCA and factor analysis is an important tool of data analysis. A critical step is selecting the number of components. However, existing methods (such as the scree plot, likelihood ratio, parallel analysis, etc) do not have statistical guarantees in the increasingly common setting where the data are heterogeneous. There each noise entry can have a different distribution. To address this problem, we propose the Signflip Parallel Analysis (Signflip PA) method: it compares data singular values to those of "empirical null" data generated by flipping the sign of each entry randomly with probability one-half. We show that Signflip PA consistently selects factors above the noise level in high-dimensional signal-plus-noise models (including spiked models and factor models) under heterogeneous settings. Here classical parallel analysis is no longer effective. To do this, we propose to leverage recent breakthroughs in random matrix theory, such as dimension-free operator norm bounds [Latala et al, 2018, Inventiones Mathematicae], and large deviations for the top eigenvalues of nonhomogeneous matrices [Husson, 2020]. To our knowledge, some of these results have not yet been used in statistics. We also illustrate that Signflip PA performs well in numerical simulations and on empirical data examples.

An Lp theory of PCA and spectral clustering

Emmanuel Abbe[1]*, Jianqing Fan*[2] *and* ♦*Kaizheng Wang*[3]
[1]EPFL
[2]Princeton University
[3]Columbia University
kw2934@columbia.edu

Principal Component Analysis (PCA) is a powerful tool in statistics and machine learning. While existing study of PCA focuses on

the recovery of principal components and their associated eigenvalues, there are few precise characterizations of individual principal component scores that yield low-dimensional embedding of samples. That hinders the analysis of various spectral methods. In this paper, we first develop an ℓ_p perturbation theory for a hollowed version of PCA in Hilbert spaces which provably improves upon the vanilla PCA in the presence of heteroscedastic noises. Through a novel ℓ_p analysis of eigenvectors, we investigate entrywise behaviors of principal component score vectors and show that they can be approximated by linear functionals of the Gram matrix in ℓ_p norm, which includes ℓ_2 and ℓ_∞ as special cases. For sub-Gaussian mixture models, the choice of p giving optimal bounds depends on the signal-to-noise ratio, which further yields optimality guarantees for spectral clustering. For contextual community detection, the ℓ_p theory leads to simple spectral algorithms that achieve the information threshold for exact recovery and the optimal misclassification rate.

Estimation of spectra of high-dimensional separable covariance matrices

♦*Lili Wang*[1] *and Debashis Paul*[2]
[1]Zhejiang Gongshang University
[2]University of California, Davis
liliwang@mail.zjgsu.edu.cn

The aim of this work is to estimate the joint spectra of high-dimensional time series for which the observed data matrix is assumed to have a separable covariance structure. The primary interest is in estimating the distribution of the eigenvalues of the marginal covariance of the observation vectors under partial information – such as stationarity or sparsity – on the temporal covariance structure. We develop a method that utilizes random matrix theory to estimate the unknown population spectra by repressing the spectrum of the dimensional covariance matrix on a simplex. We prove the consistency of the proposed estimator under the dimension proportional to the sample size setting. Furthermore, we develop a resampling based method for statistical inference on low-dimensional functionals of the joint spectrum of the population covariance matrix.

Session 99: Analyses of Electronic Health Records

Statistical inference for natural language processing algorithms when predicting type 2 diabetes using electronic health record notes

Brian Egleston[1] *and* ♦*Ashis Chanda*[2]
[1]Fox Chase Cancer Center
[2]Temple University
brian.egleston@fccc.edu

The Pointwise Mutual Information statistic (PMI), which measures how often two words occur together in a document corpus, is a cornerstone of recently proposed popular natural language processing algorithms such as word2vec. PMI and word2vec reveal semantic relationships between words and can be helpful in a range of applications such as document indexing, topic analysis, or document categorization. We use probability theory to demonstrate the relationship between PMI and word2vec. We use the theoretical results to demonstrate how the PMI can be modeled and estimated in a simple and straight forward manner. We further describe how one can obtain standard error estimates that account for within-patient clustering that arises from patterns of repeated words within a patient's health record due to a unique health history. We then demonstrate the usefulness of PMI on the problem of predictive identification of disease from free text notes of electronic health records. Specifically, we use our methods to distinguish those with and without type 2 diabetes mellitus in electronic health record free text data using over 400,000 clinical notes from an academic medical center. This is joint work with Tian Bai, Richard J. Bleicher, Stanford J. Taylor, Michael H. Lutz, and Slobodan Vucetic.

Prediction of relapse in chronic disease using fragmented medical records

Jiasheng Shi and ♦*Jing Huang*
University of Pennsylvania
jing14@upenn.edu

Most patients with chronic illnesses experience symptoms on and off throughout their life. Understanding differential remission and relapse risk profiles and predicting the risk of relapses for an individual patient is critical for optimizing disease management and improving long-term outcomes. However, medical records collected during clinical practice provide fragmented data with episodes of disease activities mostly observed after the onset of relapse. Disease activity during remission is often unobserved. In this study, we propose a method to improve the prediction of relapse using fragmented data from medical records. The method is applied to pediatric ulcerative colitis to better understand the remission and relapse cycle.

Handling irregular observation in longitudinal studies using electronic health records

Eleanor Pullenayegum
The Hospital for Sick Children
eleanor.pullenayegum@sickkids.ca

Longitudinal data collected as part of usual healthcare delivery are becoming increasingly available for research through electronic health records. However, a common feature of these data is that they are collected more frequently when patients are unwell. For example, newborns who are slow to regain their birthweight will require more frequent monitoring and will consequently have more weight measurements than their typically growing counterparts. Failing to account for this would lead to underestimation of the rate of growth of the population of newborns as a whole. I will discuss approaches to handling the informative nature of the observation, including recent developments for handling clustering of patients (e.g within hospitals). I will provide some simulation results as well as an application to the inter-relationships among air quality, wheezing, and outdoor play among children living in the Greater Toronto Area.

Statistical Methods for Phenotyping with Positive-Only Electronic Health Record Data

Jinbo Chen
University of Pennsylvania
NA

EHR phenotyping models are generally developed using data from a group of patients with known statuses of an interest condition. Due to asymmetric clinical workflow, it is often convenient to identify a group of cases. Data from these labeled cases and a large number of unlabeled patients, referred to as "positive-only" data, is then available with minimum requirement for labeling efforts. We develop statistical methods to provide methodological foundation for positive-only data to be routinely used for training and validating phenotyping models. We performed extensive simulation studies to illustrate performance of these methods and applied them to Penn Medicine EHR data to phenotype primary aldosteronism.

Session 100: Integrative multi-omics inference

Double-matched matrix decomposition for multi-view data

Dongbang Yuan and ♦*Irina Gaynanova*
Texas A&M University
Irinag@stat.tamu.edu

We consider the problem of extracting joint and individual signals from multi-view data, that is data collected from different sources on matched samples. While existing methods for multi-view data decomposition explore single matching of data by samples, we focus on double-matched multi-view data (matched by both samples and source features). Our motivating example is the miRNA data collected from both primary tumor and normal tissues of the same subjects; the measurements from two tissues are thus matched both by subjects and by miRNAs. Our proposed double-matched matrix decomposition allows to simultaneously extract joint and individual signals across subjects, as well as joint and individual signals across miRNAs. Our estimation approach takes advantage of double-matching by formulating a new type of optimization problem with explicit row space and column space constraints, for which we develop an efficient iterative algorithm. Numerical studies indicate that taking advantage of double-matching leads to superior signal estimation performance compared to existing multi-view data decomposition based on single-matching. We apply our method to miRNA data as well as data from the English Premier League soccer matches, and find joint and individual multi-view signals that align with domain specific knowledge.

Deep IDA: A Deep Learning Method for Integrative Discriminant Analysis of Multi-View Data with Feature Ranking

Jiuzhou Wang and ♦*Sandra Safo*
University of Minnesota
ssafo@umn.edu

Improvements in technologies produce an unprecedented amount of diverse but related data (e.g., genetics, proteomics) that, in addition with clinical data, can be leveraged to effectively predict an outcome and to identify important variables. Recent work show that prediction methods that leverage the strengths of these multi-faceted or multi-view data simultaneously (i.e., one-step methods) have enormous potential to yield more powerful findings than two-step methods: association followed by prediction. Most existing one-step methods have focused on linear associations, but the relationships between data from multiple views, and data from multiple views and an outcome, however, are too complicated to be understood solely by linear methods. Further, the existing nonlinear one-step method does not allow for feature selection, a key factor for having explainable models. We propose a new deep learning method, Deep IDA, for joint association and classification of data from multiple views that permits feature ranking and enhances our ability to identify features from each view that contribute to the association of the views and separation of classes within each view. Our framework for feature ranking is general and adaptable to many deep learning methods to enhance explanation of deep learning models. Simulations result demonstrate the superior classification and feature selection performance of the method in comparison with state-of-the-art methods. Application of the proposed method to omics data from patients with and without COVID-19 have identified biomolecules shedding light on the molecular architecture of COVID-19 disease and severity.

Integrating multi-omics data for causal inference

♦*Dan Zhou*[1] *and Eric Gamazon*[2]
[1]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA
[2]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA; Clare Hall, University of Cambridge, Cambridge, UK; MRC Epidemiology Unit, University of Cambridge, Cambridge, UK
dan.zhou@vumc.org

We developed a joint-tissue imputation (JTI) approach and a Mendelian randomization (MR) framework for causal inference, MR-JTI. By integrating cell-type level epigenomic signatures, JTI borrows information across transcriptomes of different tissues, leveraging shared genetic regulation, to improve prediction performance in a tissue-dependent manner. Notably, JTI includes the single-tissue imputation method PrediXcan (Gamazon et al. Nat Genet 2015) as a special case and outperforms other imputation approaches. MR-JTI models variant-level heterogeneity (primarily due to horizontal pleiotropy, addressing a major challenge of transcriptome-wide association study interpretation) and performs causal inference with type I error control. We make explicit the connection between the genetic architecture of gene expression and of complex traits and the suitability of Mendelian randomization as a causal inference strategy for transcriptome-wide association studies. Analysis of biobanks and meta-analysis data, and extensive simulations show substantially improved statistical power for the framework. The pre-trained imputation models generated from the latest Genotype-Tissue Expression (GTEx) dataset are available for public use.

Session 101: New advances of statistical inference for high-dimensional data

Power-enhanced simultaneous test of high-dimensional mean vectors and covariance matrices with application to gene-set testing

♦*Xiufan Yu*[1]*, Danning Li*[2]*, Lingzhou Xue*[3] *and Runze Li*[3]
[1]University of Notre Dame
[2]Northeast Normal University
[3]Pennsylvania State University
xzy22@psu.edu

Power-enhanced tests with high-dimensional data have received growing attention in theoretical and applied statistics in recent years. Existing tests possess their respective high-power regions, and we may lack prior knowledge about the alternatives when testing for a problem of interest in practice. There is a critical need of developing powerful testing procedures against more general alternatives. This work studies the joint test of two-sample mean vectors and covariance matrices for high-dimensional data. We first expand the high-power region of high-dimensional mean tests or covariance tests to a wider alternative space and then combine their strengths together in the simultaneous test. We develop a new power-enhanced simultaneous test that is powerful to detect differences in either mean vectors or covariance matrices under either sparse or dense alternatives. We prove that the proposed testing procedures align with the power enhancement principles introduced by Fan et al. (2015) and achieve the accurate asymptotic size and consistent asymptotic power. We demonstrate the finite-sample performance using simulation studies and a real application to find differentially expressed gene-sets in cancer studies. Our findings in the empirical study are supported by the biological literature.

A Distribution-Free Independence Test for High Dimension Data

♦*Zhanrui Cai, Jing Lei and Kathryn Roeder*
Carnegie Mellon University
zxc55@psu.edu

Test of independence is a fundamental yet difficult topic in statistics. Without distributional or structural assumptions, it is extremely difficult to perform independent testing when the data is of high dimension and the dependence signal is sparse. In this paper, we solve this problem by proposing a general framework for independence testing that borrows strength from the most powerful classification tools, such as neural networks. We implement only one fixed permutation of the data to construct a permuted sample that has the same distribution as the original data under the null hypothesis. By using sample splitting, the classifier is trained on the first subset of data, aiming to distinguish the permuted data from the original data. We then construct the test statistic using the classifier and the second half of the data. The test statistic has a limiting standard normal distribution and doesn't require the computationally expensive bootstrap to calculate the p-value. Extensive simulations are conducted to illustrate the advantages of the newly proposed test compared with previous methods. We further apply the new test to a single cell data set to test the independence of RNA-seq and ATAC-seq.

Optimal sampling designs for online estimation of streaming multi-dimensional time series

Shuyang Bai[1], ♦*Rui Xie*[2] *and Ping Ma*[1]
[1]University of Georgia
[2]University of Central Florida
Rui.Xie@ucf.edu

Online analysis of streaming time series data often faces a trade-off between statistical efficiency and computational cost. One important approach to balance this trade-off is sampling, where only a small portion of the sample is selected for the model fitting and update. In this paper, we study the data-dependent sample selection and online analysis problem for a multi-dimensional streaming time series. Motivated by the D-optimality criterion in design of experiments, we propose a class of online data reduction methods that achieve an optimal sampling criterion and improve the computational efficiency of the online analysis. We show that the optimal solution amounts to a strategy that is a mixture of Bernoulli sampling and leverage score sampling. The leverage score sampling involves an auxiliary estimate of an inverse covariance matrix that is updated sparsely to gain the computational advantage. Theoretical properties of the auxiliary estimations involved are established. The performance of the sampling-assisted online estimation method is assessed via simulation studies and a real data example.

Multiple autocovariance change-points detection for high-dimensional time series

Di Xiao[1], ♦*Haotian Xu*[2], *Yuan Ke*[1] *and Jeongyoun Ahn*[1]
[1]University of Georgia
[2]University of Warwick
Haotian.Xu@unige.ch

We consider two problems about change-points in autocovariance structures of a high-dimensional time series with heavy-tailed innovations. First, we study a multiple change-points detection method based on the matrix max norm of the tail-robust moving sum (MOMSUM) statistic. From a nonasymptotic perspective, we characterize the interplay among the window size of MOMSUM statistic, the dimensionality, the minimum spacing between consecutive change points, and the value of smallest change. While from an asymptotic perspective, under mild conditions, we show that the number and the locations of change-points 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 change-point at a prespecified location.

Session 102: Recent Advances in Analysis of Data with Measurement Errors

Methods for diagnostic accuracy with biomarker measurement error

♦*Ching-Yun Wang and Ziding Feng*
Fred Hutchinson Cancer Research Center
cywang@fredhutch.org

Diagnostic biomarkers are often measured with errors due to imperfect lab conditions or temporal variability within individuals. The ability of a diagnostic biomarker to discriminate between cases and controls is often measured by the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, among others. Ignoring measurement error can cause biased estimation of a diagnostic accuracy measure which results in misleading interpretation of the efficacy of a diagnostic biomarker. Existing assays available are either research grade or clinical grade. Research assays are cost effective, often multiplex, but they may be associated with moderate measurement errors leading to poorer diagnostic performance. In comparison, clinical assays may provide better diagnostic ability, but with higher cost since they are usually developed by industry. However, diagnostic companies often are not interested in investing until an adequate diagnostic performance is observed. Therefore, a significant challenge is to select biomarker candidates for further development when their potentials are not fully observed while only research assays with varying analytical variability are available. In this paper, we develop methods to correct bias in estimating diagnostic performance measures including AUC, sensitivity, and specificity of 2 different error–prone assay measurements. An important strength of our methods is that we do not need to assume availability of biomarker replicates or a validation subset. Our methods can be applied to evaluate the diagnostic efficacy of clinical assays in comparison with research assays. Finite sample performance of the proposed method is examined via extensive simulation studies. The methods are applied to a pancreatic cancer study.

Robust estimation of the causal risk difference with misclassified outcome data

♦*Di Shu*[1] *and Grace Yi*[2]
[1]University of Pennsylvania & Children's Hospital of Philadelphia
[2]University of Western Ontario
shudi1991@gmail.com

Misclassification frequently occurs in real-world data - including electronic health records and administrative claims data - and continues to impose a challenge to statistical analysis. In order to obtain reliable inference results, statistical analysis should take the misclassification effect into account. In this talk, we will focus on estimation of the causal risk difference in the presence of outcome misclassification. I will first introduce a closed-form, bias-corrected estimator when using inverse probability of treatment weighting. Then, I will extend the results to correct for outcome misclassification in the context of doubly robust estimation. That is, the resulting bias-adjusted estimator is consistent when the treatment model or the outcome model is correctly specified. Finally, to achieve more protection against model misspecification, I will il-

lustrate how to eliminate misclassification bias when applying the empirical likelihood-based multiply robust estimation. Our estimator corrects for misclassification while preserving the multiple robustness in that they are guaranteed to be consistent when either a set of postulated treatment models or a set of postulated outcome models contains a correctly specified model. We demonstrate the use and assess the performance of the proposed methods through an application to smoking cessation data and extensive simulations.

Addressing measurement error in random forests using quantitative bias analysis

Tammy Jiang

Boston University School of Public Health
tjiang1@bu.edu

Machine learning is increasingly used to predict health outcomes and detect novel risk factors for diseases. Although measurement error is pervasive, the impact of measurement error on machine learning predictions is seldom quantified. The purpose of this study was to assess the impact of measurement error on random forest model performance and variable importance using quantitative bias analysis. Quantitative bias analysis is an epidemiologic approach to quantifying the influence of systematic error on study results. First, we assessed the impact of misclassification (i.e., measurement error of categorical variables) of predictors on random forest model performance and variable importance (mean decrease in accuracy) using data from the United States National Comorbidity Survey Replication. Second, we simulated datasets in which we know the true model performance and variable importance measures and could verify that quantitative bias analysis recovered the truth in misclassified versions of the datasets. We found that measurement error in the data used to construct random forests can distort model performance and variable importance measures, and that quantitative bias analysis can recover the correct results. With the growing use of machine learning and availability of big data, it is crucial to address measurement error. Quantitative bias analysis is one method that can be used to quantify the impact of measurement error on machine learning results.

Corrected Score Methods for Recurrent Event Data with Covariate Measurement Error

Hsiang Yu[1], ♦*Yu-Jen Cheng*[2] *and Ching-Yun Wang*[3]

[1]Taiwan Semiconductor Manufacturing

[2]National Tsing Hua University

[3]Fred Hutchinson Cancer Research Center
ycheng@stat.nthu.edu.tw

Recurrent event data arise frequently in many longitudinal follow-up studies. Examples include repeated hospitalizations, recurrent infections in HIV, and tumor recurrences. In contrast to the existing works, in our case the conventional assumption of independent censoring is violated since the recurrent event process is interrupted by some correlated terminal events. Further, some covariates may be measured with errors. To accommodate both informative censoring and measurement error, the occurrence of recurrent events is modelled through an unspecified frailty distribution and accompanied with a classical measurement error model. In this work, corrected score approaches are developed to correct bias for the estimation of parameters in the recurrent event model even when the informative censoring occurs. The asymptotic properties of the proposed estimators are established, and the finite sample performances are examined via simulations. The proposed methods are applied to a real data.

Session 103: Modern streaming Data Analysis: anomaly detection and applications

Solar Radiation Anomaly Events Modeling Using Spatial-Temporal Mutually Interactive Processes

Minghe Zhang[1], *Chen Xu*[1], *Andy Sun*[1], *Feng Qiu*[2] *and* ♦*Yao Xie*[1]

[1]Georgia Institute of Technology

[2]Argonne National Lab
yao.xie@isye.gatech.edu

Modeling and predicting solar events, in particular, the solar ramping event is critical for improving situational awareness for solar power generation systems. Solar ramping events are significantly impacted by weather conditions such as temperature, humidity, and cloud density. Discovering the correlation between different locations and times is a highly challenging task since the system is complex and noisy. We propose a novel method to model and predict ramping events from spatial-temporal sequential solar radiation data based on a spatio-temporal interactive Bernoulli process. We demonstrate the good performance of our approach on real solar radiation datasets.

A Change-Point Marginal Regression Model for Stock Returns' Tail Exceedance under Market Variations

Haipeng Xing

Stony Brook University
haipeng.xing@stonybrook.edu

Tail behaviors of asset price changes have been noted to be closely related to assets' trad- ing history and market variations. To investigate this issue, we model intensities of tail exceedances via a marginal regression model with multiple change-points in regression co- efficients, which represents large and small changes of market environment. We develop a mixtured estimating equation approach and a segmentation procedure for regression co- efficients and unobserved change-points, and demonstrate their large sample properties. We then analyze a market consisting of 500 U.S. stocks during 2010-2013 and discuss the implication of our result to market practice.

Does enforcing fairness mitigate biases caused by subpopulation shift?

Subha Maity[1], *Debarghya Mukherjee*[1], *Mikhail Yurochkin*[2] *and* ♦*Yuekai Sun*[1]

[1]University of Michigan

[2]MIT-IBM Watson AI Lab
yuekai@umich.edu

Many instances of algorithmic bias are caused by subpopulation shifts. In this paper, we study whether enforcing algorithmic fairness during training mitigates such biases in the target domain. On one hand, we show that enforcing fairness may not improve performance on minority groups (in the target domain). In fact, it may even harm performance on minority groups. On the other hand, we derive necessary and sufficient conditions under which enforcing algorithmic fairness totally mitigates biases due to subpopulation shifts. We also illustrate the practical implications of our theoretical results in simulations and on real data.

Homeostasis phenomenon in conformal prediction and predictive distribution functions

Minge Xie

Rutgers University
mxie@stat.rutgers.edu

Conformal prediction is an attractive framework for prediction that is error distribution free in machine learning and statistics.In this

article, we study in details its "homeostasis property" under a general regression setup and also introduce the concepts of upper and lower predictive distributions and predictive curve to establish connections to left-, right- and two-tailed hypothesis testing problems as well as the developments in confidence distributions. The homeostasis property is very attractive, since it states that under some conditions the prediction results remain valid even if the model used for learning is completely wrong. We show explicitly why the property holds in a model-based setup and also explore the boundary when the property breaks down. Beside the typical assumption used in conformal prediction that the response and covariate pairs (y, x) of all subjects are iid distributed, we also study the classical regression setting in which the design is fixed with given (non-random) covariates. The trade-offs among learning model accuracy, prediction valid and prediction efficiency are discussed, leading to an emphasis of more efforts on developing better learning models. The investigation provides an explanation why a "black box" approach (such as a deep neural network method) works well for prediction in many academic exercises but meets with more mixed results in real world applications.

Session 104: Enhancing Causal Inference Using Genetics Data: Mendelian Randomization

Exploiting public GWAS databases to identify and adjust for heritable confounders in Mendelian randomization studies.

Jean Morrison
University of Michigan
jvmorr@umich.edu

Mendelian randomization (MR) is a widely used causal inference method which promises tests of causal effects derived only from genetic associations. When applied carefully, MR can yield valuable insight. However, pleiotropy and heritable confounding remain common sources of MR false positives. The most reliable MR studies rely on pre-existing evidence and expertise to identify and control for possible sources of confounding. However, for many promising applications of MR, little prior information is available. In lieu of such information, we propose an automated method that exploits the deep and growing resources of publicly available GWAS summary statistics. Our method identifies clusters of pleiotropic loci which may have effects mediated by a heritable confounder. The method returns to the investigator a list of potential confounders as well as adjusted estimates of the causal effect. The investigator can then apply their domain knowledge to accept or reject potential confounders. This strategy can serve a dual role of providing more robust MR results and suggesting new candidates for further investigation.

Mendelian Randomization Analysis Using Multiple Biomarkers of an Underlying Common Exposure

♦*Jin Jin*[1]*, Guanghao Qi*[2]*, Zhi Yu*[3] *and Nilanjan Chatterjee*[1]
[1]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University
[2]Department of Biomedical Engineering, Johns Hopkins University
[3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard
jjin31@jhu.edu

Mendelian Randomization (MR) analysis is increasingly popular for testing the causal effect of exposures on disease outcomes using data from genome-wide association studies. In some settings, the underlying exposure, such as systematic inflammation, may not be directly observable, but measurements can be available on multiple biomarkers, or other types of traits, that are co-regulated by the exposure. We propose a method for MR analysis with Latent Exposure (MRLE), which tests the significance for, and the direction of, the effect of a latent exposure by leveraging information from multiple related traits. The method is developed by constructing a set of estimating functions based on the second-order moments of GWAS summary association statistics for the observable traits, under a structural equation model where genetic variants are assumed to have indirect effects through the latent exposure and potentially direct effects on the traits. Simulation studies showed that MRLE has well-controlled type I error rates and enhanced power compared to single-trait MR tests under various types of pleiotropy. Applications of MRLE using genetic association statistics across five inflammatory biomarkers (CRP, IL-6, IL-8, TNF-alpha and MCP-1) provided evidence for potential causal effects of inflammation on increasing the risk of coronary artery disease, colorectal cancer and rheumatoid arthritis, while standard MR analysis for individual biomarkers failed to detect consistent evidence for such effects.

Almost exact Mendelian randomization

♦*Matthew Tudball*[1] *and Qingyuan Zhao*[2]

[1]University of Bristol

[2]University of Cambridge

matt.tudball@bristol.ac.uk

Mendelian randomization (MR) is an observational design that bases inference on the random transmission of alleles from parents to offspring. However, this inferential basis is typically only implicit in MR methodologies or used as an informal justification. As parent-offspring data becomes more widely available, we advocate an approach to MR which is exactly based on this randomization, making explicit the common analogy between MR and a randomized controlled trial. We begin by developing a complete causal framework for MR which connects and formalizes phenomena such as population structure, gamete formation, fertilization, genetic linkage and pleiotropy. We use this framework to detect biases in the MR design and identify sufficient confounder adjustment sets to correct them. We then introduce a hypothetical idealized inference for MR based on exact hypothesis testing, first described in RA Fisher's original proposal for randomized experiments. Exact inference requires knowledge of the distribution of offspring haplotypes resulting from meioses in one or both parents. Although this knowledge is not available, meiosis has been thoroughly studied and modelled in genetics dating back to Haldane (1919). We can therefore perform almost exact inference by substituting a reasonable model-based approximation for the randomization distribution. While randomization-based inference offers a transparent and conceptually appealing approach to MR, it also offers some practical advantages. First, unlike existing within-family MR methods, it sidesteps the need for correctly specifying phenotype models, although a better model will often lead to more powerful tests. We demonstrate via simulation that propensity scores obtained from the underlying meiosis model can form powerful test statistics. Second, our approach is robust to arbitrarily weak instruments. Finally, by using our sufficient adjustment sets, it is provably robust to biases arising from population structure, assortative mating, dynastic effects and pleiotropy via linkage disequilibrium.

Session 105: Recent Development in Functional Data Analysis

In-game win probabilities for the National Rugby League

♦*Tianyu Guan*[1]*, Robert Nguyen*[2]*, Jiguo Cao*[3] *and Tim Swartz*[3]
[1]Brock University
[2]University of New South Wales
[3]Simon Fraser University
tguan@brocku.ca

We develop new methods for providing instantaneous in-game win probabilities for the National Rugby League. Besides the score differential, betting odds and real-time features extracted from the match event data are also used as inputs to inform the in-game win probabilities. Rugby matches evolve continuously in time and the circumstances change over the duration of the match. Therefore, the match data are considered as functional data, and the in-game win probability is a function of the time of the match. We express the in-game win probability using a conditional probability formulation, the components of which are evaluated from the perspective of functional data analysis. Specifically, we model the score differential process and functional feature extracted from the match event data as sums of mean functions and noises. The mean functions are approximated by B-spline basis expansions with functional parameters. Since each match is conditional on a unique kickoff win probability of the home team obtained from the betting odds (i.e. the functional data are not independent and identically distributed), we propose a weighted least squares method to estimate the functional parameters by borrowing the information from matches with similar kickoff win probabilities. The variance and covariance elements are obtained by the maximum likelihood estimation method. The proposed method is applicable to other sports when suitable match event data are available.

Temporal-dependent Principal Component Analysis of Two-dimensional Functional Data

♦*Kejun He*[1]*, Shirun Shen*[1] *and Lan Zhou*[2]
[1]Renmin University of China
[2]Texas A&M University
kejunhe@ruc.edu.cn

In this work, we propose a novel model to analyze the temporal-dependent two-dimensional functional data on an irregular domain. Illustrated by a study of Texas temperature, our method assumes that the functional principal component scores of two-dimensional functional data are temporally correlated and models the scores as latent time series. To overcome the challenge that the two-dimensional functions of interest are irregularly and sparsely observed, we use the bivariate spline basis on triangulations. All of these ideas are integrated into a unified model and an expectation-maximization algorithm along with Kalman filter and smoother is developed to estimate the unknown parameters. A simulation study is conducted to demonstrate that the proposed model outperforms its alternative. We finally use the proposed model to analyze the dataset of Texas temperature and the results are consistent with the scientific conclusions in domain knowledge.

Inference on Linear Models for Functional Data via Bootstrapping Max Statistics

Yinan Lin
National Univerisity of Singapore
stayina@nus.edu.sg

We develop a unified approach for hypothesis testing in various types of functional linear models, such as scalar-on-function, function-on-function, function-on-scalar models, that have a wide range of applications in functional data analysis. In addition, the proposed test can handle models of mixed types, for example, models with both functional and scalar/vector predictors. Unlike most existing methods that rest on large-sample distributions of test statistics, the proposed method leverages the technique of bootstrapping max statistics and exploits variance decay, an inherent feature of functional data, to achieve excellent numerical performance even when the sample size is limited.

Basis Expansions for Functional Snippets

Zhenhua Lin[1]*, Jane-Ling Wang*[2] *and* ♦*Qixian Zhong*[3]
[1]National University of Singapor
[2]University of California
[3]Xiamen University
zhqx016@163.com

Estimation of mean and covariance functions is fundamental for functional data analysis. While this topic has been studied extensively in the literature, a key assumption is that there are enough data in the domain of interest to estimate both the mean and covariance functions. In this paper, we investigate mean and covariance estimation for functional snippets in which observations from a subject are available only in an interval of length strictly (and often much) shorter than the length of the whole interval of interest. For such a sampling plan, no data is available for direct estimation of the off-diagonal region of the covariance function. We tackle this challenge via a basis representation of the covariance function. The proposed estimator en- joys a convergence rate that is adaptive to the smoothness of the underlying covariance function, and has superior finite-sample performance in simulation studies.

Session 106: New Advances in High-Dimensional Time Series Analysis

Learning Financial Network with Focally Sparse Structure

♦*Weining Wang*[1]*, Chen Huang*[2] *and Victor Chernozhukov*[3]
[1]U of York
[2]U of aarhus
[3]MIT
weining.wang@york.ac.uk

This paper studies the estimation of network connectedness with focally sparse structure. We try to uncover the network effect with a flexible sparse deviation from a predetermined adjacency matrix. To be more specific,the sparse deviation structure can be regarded as latent or misspecified linkages. To obtain high-quality estimator for parameters of interest,we propose to use a double regularized high-dimensional generalized method of moments (GMM) framework. Moreover,this framework also facilitates us to conduct the inference. The oretical results on consistency and asymptotic normality are pro vided with accounting for general spatial and temporal dependency of the underlying data generating processes. Simulations demonstrate good performance of our proposed procedure. Finally,we apply the methodology to study the spatial network effect of stock returns.

Online Inference with Stochastic Gradient Descent via Random Scaling

Simon Lee[1]*, Matt Seo*[2]*, Youngki Shin*[3] *and* ♦*Yuan Liao*[4]
[1]Columbia
[2]Seoul national university
[3]McMaster University
[4]Rutgers
yuan.liao@rutgers.edu

We develop a new online inference method for parameters estimated by the Polyak-Ruppert averaging of stochastic gradient descent (SGD) algorithms. We leverage insights from time series regression in econometrics and construct asymptotically pivotal statistics via random scaling. Our approach is fully operational with online data and is rigorously underpinned by a functional central limit theorem. The test statistic is computed in an online fashion with only SGD iterates and the critical values can be obtained with no need to estimate the asymptotic variance.

Title: Forecasting time series with Wasserstein GANs

Moritz Haas and ♦*Stefan Richter*
Heidelberg University
stefan.richter@iwr.uni-heidelberg.de

In this talk, I will present some theoretical contributions on forecasting high-dimensional time series with Wasserstein generative adversarial networks. We use them to estimate the conditional distribution of the next observation given the past which allows for confidence intervals. The behavior of the estimates is analyzed in an application with temperature data. Under structural conditions and smoothness assumptions on the underlying evolution of the time series, we prove convergence rates which do not suffer from the curse of dimension. To do so, we use and provide new results in empirical process theory for dependent data.

Session 107: Recent Developments for network data analysis: Theory, Method, and Application

Euclidean Representation of Low-Rank Matrices and Its Statistical Applications

Fangzheng Xie
Indiana University
fxie@iu.edu

Low-rank matrices are pervasive throughout statistics, machine learning, signal processing, optimization, and applied mathematics. In this paper, we propose a novel and user-friendly Euclidean representation framework for low-rank matrices. Correspondingly, we establish a collection of technical and theoretical tools for analyzing the intrinsic perturbation of low-rank matrices in which the underlying referential matrix and the perturbed matrix both live on the same low-rank matrix manifold. Our analyses show that, locally around the referential matrix, the sine-theta distance between subspaces is equivalent to the Euclidean distance between two appropriately selected orthonormal basis, circumventing the orthogonal Procrustes analysis. We also establish the regularity of the proposed Euclidean representation function, which has a profound statistical impact and a meaningful geometric interpretation. These technical devices are applicable to a broad range of statistical problems. Specific applications considered in detail include Bayesian sparse spiked covariance model with non-intrinsic loss, efficient estimation in stochastic block models where the block probability matrix may be degenerate, and least-squares estimation in biclustering problems. Both the intrinsic perturbation analysis of low-rank matrices and the regularity theorem may be of independent interest.

Block Mean-mean-field variational inference for dynamic latent space models

♦*Peng Zhao, Debdeep Pati, Anirban Bhattacharya and Bani Mallick*
Texas A&M University
pzhao@tamu.edu

We consider a latent space model for dynamic networks in this paper. The objective is to estimate all the pairwise inner products of the latent positions. To balance the posterior inference and computational scalability, we propose a block mean-field variational inference framework, where the dependence across time of the dynamic networks is exploited to facilitate the computation and inference. In addition, a simple-to-implement block coordinate ascent algorithm with message-passing type of updatings in each block is developed. To explore the frequentist property of the posterior distribution, we assume the ℓ_1 norm of the total variation across time of true latent positions is bounded. We first show the lower bound to the minimax risk under the given parameter space. Moreover, we show that both the fractional posterior and the variational risk for our proposed variational inference approach under frequently used Gaussian random walk priors attains the rate of minimax lower bounds under certain conditions. Finally, simulations and real data analysis demonstrate the efficacy of our methodology and the efficiency of our algorithm.

Network Functional Varying Coefficient Model

Xuening Zhu
Fudan University
xueningzhu@fudan.edu.cn

We consider functional responses with network dependence observed for each individual at irregular time points. To model both the inter-individual dependence and within-individual dynamic correlation, we propose a network functional varying coefficient (NFVC) model. The response of each individual is characterized by a linear combination of responses from its connected nodes and its exogenous covariates. All the model coefficients are allowed to be time dependent. The NFVC model adds to the richness of both the classical network autoregression model and the functional regression models. To overcome the complexity caused by the network interdependence, we devise a special nonparametric least-squares type estimator, which is feasible when the responses are observed at irregular time points for di¡U+FB00¿erent individuals. The estimator takes advantage of the sparsity of the network structure to reduce the computational burden. To further conduct the functional principal component analysis, a novel within-individual covariance function estimation method is proposed and studied. Theoretical properties of our estimators, which involve techniques related to empirical processes, nonparametrics, functional data analysis and various concentration inequalities, are analyzed. We analyze a social network dataset to illustrate the powerfulness of the proposed procedure.

Finite Mixtures of ERGMs for Modeling Ensembles of Networks

♦*Fan Yin*[1]*, Weining Shen*[2] *and Carter Butts*[2]
[1]Microsoft
[2]University of California, Irvine
yinf2@uci.edu

Ensembles of networks arise in many scientific fields, but there are few statistical tools for inferring their generative processes, particularly in the presence of both dyadic dependence and cross-graph heterogeneity. To address this gap, we propose characterizing network ensembles via finite mixtures of exponential family random graph models, a framework for parametric statistical modeling of graphs that has been successful in explicitly modeling the complex stochastic processes that govern the structure of edges in a network. Our proposed modeling framework can also be used for applications such as model-based clustering of ensembles of networks and density estimation for complex graph distributions. We develop a joint approach to estimate the number of mixture components and iden-

tify cluster-specific parameters simultaneously as well as to obtain an identified model under the Bayesian framework. Specifically, we propose a Metropolis-within-Gibbs algorithm to perform Bayesian inference, and estimate the number of mixture components using a strategy of deliberate overfitting with sparse priors that removes excess components during MCMC. As the true ERGM likelihood is generally intractable for model specifications with dyadic dependence terms, we consider two tractable approximations (pseudolikelihood and adjusted pseduolikelihood) to facilitate efficient statistical inference. We run simulation studies to compare the performance of these two approximations with respect to multiple metrics, showing conditions under which both are useful. We also demonstrate the utility of the proposed approach using an ensemble of political co-voting networks among U.S. Senators and an ensemble of brain functional connectivity networks.

Session 108: Recent advances in Bayesian methodology for complex data

Data integration using hierarchical Gaussian process models under shape constraints

♦*Shuang Zhou*[1]*, Debdeep Pati*[2] *and Anirban Bhattacharya*[2]
[1]Arizona State University
[2]Texas A&M University
szhou98@asu.edu

Data integration has been a hot topic in real-world applications that combines data residing at different sources and extracts the shared information across sources. Hierarchical Bayesian models are a powerful tool for modeling grouped data by modeling the data and their interaction across the groups via hierarchies. We develop a method for data integration under multiple constraints using a hierarchical constrained regression with basis expansion approach. At the global level, we can incorporate multiple constraints simultaneously by finding a one-to-one mapping of the constraints on the coefficient space from the original function space under a suitable basis. At the group level, we integrate a multiplicative random effect into the sub-models with the knowledge of data annotation to estimate the group-wise unknown response deviation. We apply our model to the proton radius puzzle problem in nuclear physics, where the constraints come from the law of physics and the unknown experimental errors associate with the data sources. We recover both the global parameters and the group errors related to the sources in the synthetic data analysis and in the real application we provide reliable analyses to reconcile with the new results for the proton radius extraction.

Bayesian modeling of spatial molecular profiling data

Qiwei Li
The University of Texas at Dallas
Qiwei.Li@UTDallas.edu

The location, timing, and abundance of gene expression within a tissue define the molecular mechanisms of cell functions. Recent technology breakthroughs in spatial molecular profiling, including imaging-based technologies and sequencing-based technologies, have enabled the comprehensive molecular characterization of single cells while preserving their spatial and morphological contexts. This new bioinformatics scenario calls for effective and robust computational methods to identify genes with spatial patterns. We represent two novel Bayesian hierarchical models to analyze spatial molecular profiling data, with several unique characteristics. The first model is based on Gaussian process. It directly models the zero-inflated and over-dispersed counts. The second model is based on Ising model. It uses the energy interaction parameter to characterize a denoised spatial pattern. The Bayesian inference framework allows us to borrow strength in parameter estimation in a de novo fashion. As a result, the two proposed models show competitive performances in accuracy and robustness over existing methods in both simulation studies and two real data applications.

A large-scale Bayesian spatial extremes modeling approach with application to wind extremes in Saudi Arabia

♦*Wanfang Chen*[1]*, Stefano Castruccio*[2] *and Marc Genton*[3]
[1]East China Normal University
[2]University of Notre Dame
[3]King Abdullah University of Science and Technology
wfchen@fem.ecnu.edu.cn

Saudi Arabia has been seeking to reduce its dependence on oil by diversifying its energy portfolio, including the largely underused energy potential from wind. However, extreme winds can possibly disrupt the wind turbine operations, thus preventing the stable and continuous production of wind energy. In this study, we assess the risk of disruptions of wind turbine operations, based on return levels with a hierarchical spatial extreme modeling approach for wind speeds in Saudi Arabia. Using a unique Weather Research and Forecasting dataset, we provide the first high-resolution risk assessment of wind extremes under spatial non-stationarity over the country. We account for the spatial dependence with a multivariate intrinsic autoregressive prior at the latent Gaussian process level. The computational efficiency is greatly improved by parallel computing on subregions from spatial clustering, and the maps are smoothed by fitting the model to cluster neighbors. Under the Bayesian hierarchical framework, we measure the uncertainty of return levels from the posterior Markov chain Monto Carlo samples, and produce probability maps of return levels exceeding the cut-out wind speed of wind turbines within their lifetime. The probability maps show that locations in the South of Saudi Arabia and near the Red Sea and the Persian Gulf are at very high risk of disruption of wind turbine operations.

Dimension-free mixing for high-dimensional Bayesian variable selection

♦*Quan Zhou*[1]*, Jun Yang*[2]*, Dootika Vats*[3]*, Gareth Roberts*[4] *and Jeffrey Rosenthal*[5]
[1]Texas A&M University
[2]Oxford University
[3]Indian Institute of Technology Kanpur
[4]University of Warwick
[5]University of Toronto
quan@stat.tamu.edu

Yang et al. (Ann. Stat., 2016) proved that the symmetric random walk Metropolis-Hastings algorithm for Bayesian variable selection is rapidly mixing under mild high-dimensional assumptions. We propose a novel MCMC sampler using an informed proposal scheme, which we prove achieves a much faster mixing time that is independent of the number of covariates, under the assumptions of Yang et al. (2016). To the best of our knowledge, this is the first high-dimensional result which rigorously shows that the mixing rate of informed MCMC methods can be fast enough to offset the computational cost of local posterior evaluation. Motivated by the theoretical analysis of our sampler, we further propose a new approach called "two-stage drift condition" to studying convergence rates of Markov chains on general state spaces, which can be useful for obtaining tight complexity bounds in high-dimensional settings.

The practical advantages of our algorithm are illustrated by both simulation studies and real data analysis.

Session 109: Recent Advances in Linear Mixed Models and Applications

Informative g-priors for Mixed Models

Yu-Fang Chien[1], ♦*Haiming Zhou*[1], *Timothy Hanson*[2] *and Ted Lystig*[2]
[1]Northern Illinois University
[2]Medtronic
zhouh@niu.edu

Zellner's objective g-prior has been widely used in linear regression models due to its simple interpretation and computational tractability in evaluating marginal likelihoods. However, the g-prior further allows portioning the prior variability explained by the linear predictor versus that of pure noise. In this paper we propose a novel, yet remarkably simple g-prior specification when a subject-matter expert has information on the marginal distribution of the response. The approach is extended for use in mixed models with some surprising, but intuitive results. We compare this formulation of the g-prior with other approaches via simulation studies.

Bayesian quantile semiparametric mixed-effects double regression models

♦*Min Wang*[1], *Duo Zhang*[2], *Liucang Wu*[3] *and Keying Ye*[1]
[1]University of Texas at San Antonio
[2]Michigan Tech University
[3]Kunming University of Science and Technology
min.wang3@utsa.edu

Semiparametric mixed-effects double regression models have been used for analysis of longitudinal data in a variety of applications, as they allow researchers to jointly model the mean and variance of the mixed-effects as a function of predictors. However, these models are commonly estimated based on the normality assumption for the errors and the results may thus be sensitive to outliers and/or heavy-tailed data. Quantile regression is an ideal alternative to deal with these problems, as it is insensitive to heteroscedasticity and outliers and can make statistical analysis more robust. In this paper, we consider Bayesian quantile regression analysis for semiparametric mixed-effects double regression models based on the asymmetric Laplace distribution for the errors. We construct a Bayesian hierarchical model and then develop an efficient Markov chain Monte Carlo sampling algorithm to generate posterior samples from the full posterior distributions to conduct the posterior inference. The performance of the proposed procedure is evaluated through simulation studies and a real data application.

Small area estimation with subgroup analysis

♦*Xin Wang*[1] *and Zhengyuan Zhu*[2]
[1]Miami University
[2]Iowa State University
wangx172@miamioh.edu

In this work, a new unit level model based on a pairwise penalized regression approach is proposed for problems in small area estimation (SAE). Instead of assuming common regression coefficients for all small domains in the traditional model, the new estimator is based on a subgroup regression model which allows different regression coefficients in different groups. The alternating direction method of multipliers (ADMM) algorithm is used to find subgroups with different regression coefficients. We also consider pairwise spatial weights for spatial areal data. In the simulation study, we compare the performances of the new estimator with the traditional small area estimator. We also apply the new estimator to urban area estimation using data from the National Resources Inventory survey in Iowa.

Selecting Mixed Effects Models using Penalized Profile REML with Application to the Cohort Study of HIV

Juming Pan
42 WATSON DR
pan@rowan.edu

Selecting an appropriate structure for a linear mixed model serves as an appealing problem in a number of applications such as in the modeling of longitudinal or clustered data. In this paper, we propose a variable selection procedure for simultaneously selecting and estimating the fixed and random effects. More specifically, a profile restricted maximum likelihood (REML) function, along with an adaptive penalty, is utilized for sparse selection. The Newton-Raphson optimization algorithm is performed to complete the parameter estimation. We prove that the proposed procedure enjoys the model selection consistency. A simulation study and a cohort study of HIV application are conducted for demonstrating the effectiveness of the proposed method.

Index of Authors

See you in 2022 ICSA Applied Statistics Symposium

www.icsa.org

International Chinese Statistical Association

泛華統計協會