JOINT MEETINGS OF 2025
TAIPEI INTERNATIONAL STATISTICAL SYMPOSIUM AND 13TH ICSA INTERNATIONAL CONFERENCE

JOINT 2025

DEC 17~20

PROGRAM BOOK

Content

Welcome

On behalf of the co-chairs, we are pleased to welcome you to the 2025 Joint Meeting of the Taipei International Statistical Symposium and the 13th ICSA International Conference (Joint2025), held in Taipei, Taiwan, from December 17 to 20, 2025.

With the theme "The New Frontier: Statistical Science in a Changing World," Joint2025 brings together more than 500 participants from around the world. The conference is co-organized by the Institute of Statistical Science, Academia Sinica (ISSAS) and the International Chinese Statistical Association (ICSA), in collaboration with the Chinese Institute of Probability and Statistics (CIPS), providing an international forum for advancing statistical and data sciences in a rapidly evolving world.

The scientific program features four keynote speeches by distinguished scholars and the prestigious Pau-Lu Hsu Award Lecture. In addition, the conference includes nearly 100 invited oral sessions spanning a broad range of topics, as well as one contributed poster session highlighting emerging research and encouraging scholarly exchange.

Joint2025 is held at a particularly meaningful time as Academia Sinica approaches its 100th anniversary. Your participation adds special significance to this milestone and strengthens the impact of the conference. Alongside the academic program, we hope the social events and the vibrant setting of Taipei will foster fruitful discussions, new collaborations, and lasting professional connections.

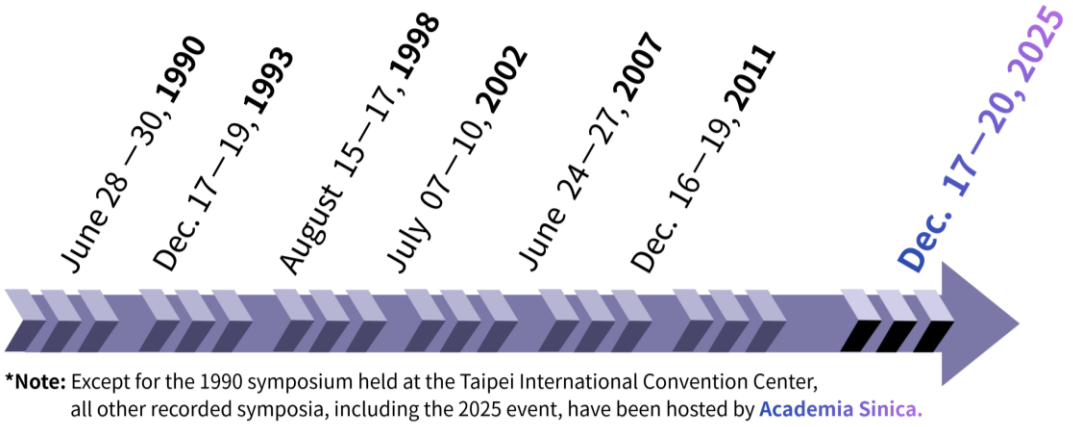
We sincerely thank all speakers, organizers, and participants for their support. We warmly welcome you to Joint2025 and wish you a productive and enjoyable conference experience in Taipei.

Co-Chairs of Joint2025
Dr. *Ming-Chung Chang*, Academia Sinica
Dr. *Xinping Cui*, University of California, Riverside
Dr. *Ying Zhang*, University of Nebraska Medical Center

Taipei International Statistical Symposium History

The Taipei International Statistical Symposium hosted by the Institute of Statistical Science, Academia Sinica is the largest statistical conference in Taiwan, ROC. It takes place about every three to five years. The origin of Taipei International Statistical Symposium dates back to 1985, since then a serial of Taipei symposiums was held in 1990, 1993, 1998, 2002, 2007, and 2011. It is occasionally jointed with different international statistical associations, such as the International Chinese Statistical Association (ICSA, 2007), Bernoulli Society (2002), and International Association for Statistical Computing - Asian Regional Section (IASC-ARS, 2011). In the 2011 meeting, there were more than 500 participants from all over the world. The Taipei International Statistical Symposium has been one of the most prestigious statistical conferences in Asian countries.

Past Taipei International Statistical Symposiums



Note:** Except for the 1990 symposium held at the Taipei International Convention Center, all other recorded symposia, including the 2025 event, have been hosted by **Academia Sinica.

Organizers and Sponsors of 2025

Main Organizer

Institute of Statistical Science, Academia Sinica (ISSAS)
International Chinese Statistical Association (ICSA)
The Chinese Institute of Probability and Statistics (CIPS)

Co-Organizer

National Science and Technology Council (NSCT)
Science Promotion and Engagement Center (SPEC)

Institute of Statistical Science Academia Sinica

The Institute of Statistical Science at Academia Sinica (ISSAS) was founded after the 14th Convocation of Academicians of Academia Sinica in July 1980, calling for its establishment. The Preparatory Office was set up in July 1982, with Dr. Min-Te Chao as the Director. The Institute was formally established on August 3, 1987, with Dr. Chao serving as the first Director. Since then, the institute has been led by a series of distinguished statisticians, with Dr. Hsin-Chou Yang taking the position in July 2023. The Institute conducts research across a wide range of areas in statistics and probability, including bioinformatics, medical and genetic statistics, brain imaging, time series, experimental design, and AI. It encourages both independent and collaborative research efforts and is known for its multidisciplinary projects. Currently, the institute has 32 research fellows, 21 postdoctoral fellows, 48 research assistants, and 21 support staff members. Over the past three years, the institute has published 326 articles in SCI journals. Its international journal, Statistica Sinica, has gained recognition as a leading statistical publication globally.

Director: Dr. Hsin-Chou Yang
Official Website: https://www.stat.sinica.edu.tw/

International Chinese Statistical Association (ICSA)

The International Chinese Statistical Association (ICSA) was established in 1991. The core mission of this professional organization is to promote the development and application of statistics within the global Chinese community. ICSA was founded to create a platform for communication, enabling Chinese statisticians, data analysts, and data scientists from different regions worldwide to share experiences, exchange ideas, and collaborate on academic research. ICSA actively organizes various academic conferences, seminars, and educational training activities aimed at promoting the latest developments and technological applications in statistics. Additionally, the association publishes professional journals, such as Statistica Sinica, co-published with the Institute of Statistical Science at Academia Sinica. These journals provide an important academic platform for statisticians to present innovative research. ICSA plays a critical role in facilitating knowledge exchange and professional growth in statistics globally, particularly within the Chinese academic and professional community. Through its activities and publications, ICSA is dedicated to supporting the education and research of statistics and fostering the professional development and collaboration of Chinese statisticians worldwide.

President: Dr. Hongyu Zhao
Official Website: https://www.icsa.org/

Committees and Chairs

Co-Chairs:

Ming-Chung Chang	Academia Sinica, Taiwan, ROC
Xinping Cui	University of California, Riverside, USA
Ying Zhang	University of Nebraska Medical Center, USA

Scientific Program Committee (ISSAS):

Chun-Shu Chen	National Central University, Taiwan, ROC
Takeshi Emura	Hiroshima University, Japan
Hsin-Cheng Huang	Academia Sinica, Taiwan, ROC
Su-Yun Huang	Academia Sinica, Taiwan, ROC
Yen-Tsung Huang	Academia Sinica, Taiwan, ROC
Ying Hung	Rutgers University, USA
Ching-Kang Ing	National Tsing Hua University, Taiwan, ROC
Mei-Ling Ting Lee	University of Maryland, USA
Fan Li	Duke University, USA
C. Jason Liang	National Institute of Allergy and Infectious Diseases, USA
Feng-Chang Lin	University of North Carolina at Chapel Hill, USA
Henry Horng-Shing Lu	National Yang Ming Chiao Tung University, Taiwan, ROC
George Michailidis	University of California, Los Angeles, USA
Shuhei Ota	Kanagawa University, Japan
Xiaotong T. Shen	University of Minnesota, USA
John Stufken	George Mason University, USA
Guei-Feng Tsai	Center for Drug Evaluation, Taiwan, ROC
Hua Tang	Stanford University, USA
Naitee Ting	Boehringer Ingelheim, USA
I-Ping Tu	Academia Sinica, Taiwan, ROC
Huixia Judy Wang	Rice University, USA
Jane-Ling Wang	University of California, Davis, USA
Hongquan Xu	University of California, Los Angeles, USA
Hsin-Chou Yang	Academia Sinica, Taiwan, ROC
Hao Zhang	Michigan State University, USA
Nancy R. Zhang	University of Pennsylvania, USA
Tingting Zhang	University of Pittsburgh, USA

Scientific Program Committee (ICSA):

Nicolas Brunel	ENSIIE & University Paris-Saclay, France
Hongyuan Cao	Florida State University, USA
Xinping Cui	University of California, Riverside, USA
Xiaowu Dai	University of California, Los Angeles, USA
Ying Ding	University of Pittsburgh, USA
Yingying Fan	University of Southern California, USA
Jesús López Fidalgo	University of Navarra, Spain
Haoda Fu	Amgen, USA
Andrew Holbrook	University of California, Los Angeles, USA
Haiyan Huang	University of California, Berkeley, USA
Jian Huang	The Hong Kong Polytechnic University, Hong Kong
Yi Li	University of Michigan, USA
Lei Liu	Washington University in St. Louis, USA
Ying Lu	Stanford University, USA
Shuangge Ma	Yale University, USA

Shujie Ma	University of California, Riverside, USA
Peihua Qiu	University of Florida, USA
Peter Song	University of Michigan, USA
Tony Sun	University of Missouri, USA
Yuanjia Wang	Columbia University, USA
Weng Kee Wong	University of California, Los Angeles, USA
Jingyuan Yang	AbbVie, USA
Emma Zhang	Emory University, USA
Xingqiu Zhao	The Hong Kong Polytechnic University, Hong Kong
Yichuan Zhao	Georgia State University, USA
Cheng Zheng	University of Nebraska Medical Center, USA

Local Organizing Committee:

Ming-Chung Chang	*Co-Chair*, Academia Sinica, Taiwan, ROC
Yi-Ju Lee	*Co-Chair*, Academia Sinica, Taiwan, ROC
Hsin-Wen Chang	Academia Sinica, Taiwan, ROC
Hsuan-Yu Chen	Academia Sinica, Taiwan, ROC
Ting-Li Chen	Academia Sinica, Taiwan, ROC
Chien-Ming Chi	Academia Sinica, Taiwan, ROC
Hsueh-Han Huang	Academia Sinica, Taiwan, ROC
Ming-Yueh Huang	Academia Sinica, Taiwan, ROC
Junho Yang	Academia Sinica, Taiwan, ROC
Chen-Hsiang Yeang	Academia Sinica, Taiwan, ROC
Tso-Jung Yen	Academia Sinica, Taiwan, ROC

General Information

Conference Venue

The conference will be held at the Building of Humanities and Social Sciences (HSSB) and the Activities Center (AC) in Academia Sinica.

Registration and Information Desk

- The registration desk and program/badge collection are located on the 3rd floor of the Humanities and Social Sciences Building (HSSB).
- Information desks are located on the 2nd floor of the AC.

Wifi Access

Academia Sinica provides the "eduroam" and "AS_Guest" wireless networks in public areas. To connect, you can either select eduroam directly, or select AS_Guest and open your browser to accept the terms of use and complete email verification. Wi-Fi is also available in conference rooms. Please refer to in-room information for details.

Note: Using the wireless network indicates acceptance of the Guest Wireless Network Service Policy.

Conference Group Photo

The conference group photo will be taken at the Humanities and Social Sciences Building at 09:00 on December 17.

Official Website: https://www3.stat.sinica.edu.tw/joint2025/

Notice to Participants

- We **DO NOT** tolerate any sexual harassment or discriminatory practices.
- Smoking is **PROHIBITED**, except in the designated smoking areas.
- All mobile phones and electronic devices should be switched to silent mode or turned off before attending any sessions or meetings.
- Should an emergency or issue arise with any equipment, please contact a Conference Assistant promptly.
- As part of our commitment to sustainability, we encourage all attendees to bring their own utensils and reusable cups. Your support for this eco-friendly initiative is highly appreciated.

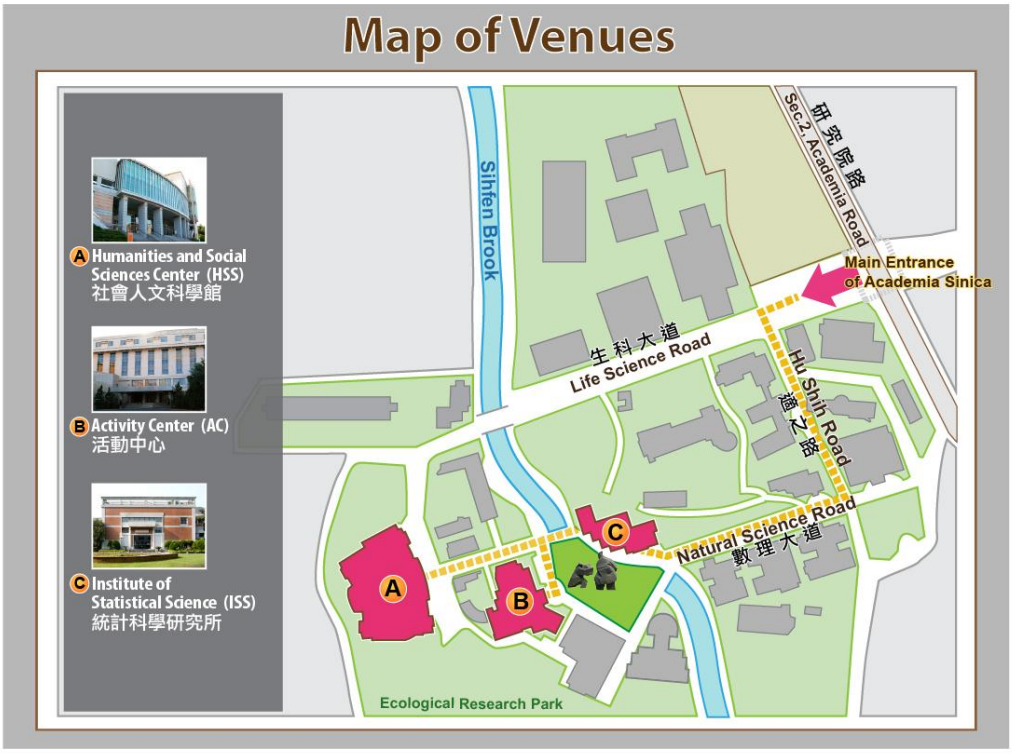
Campus Map

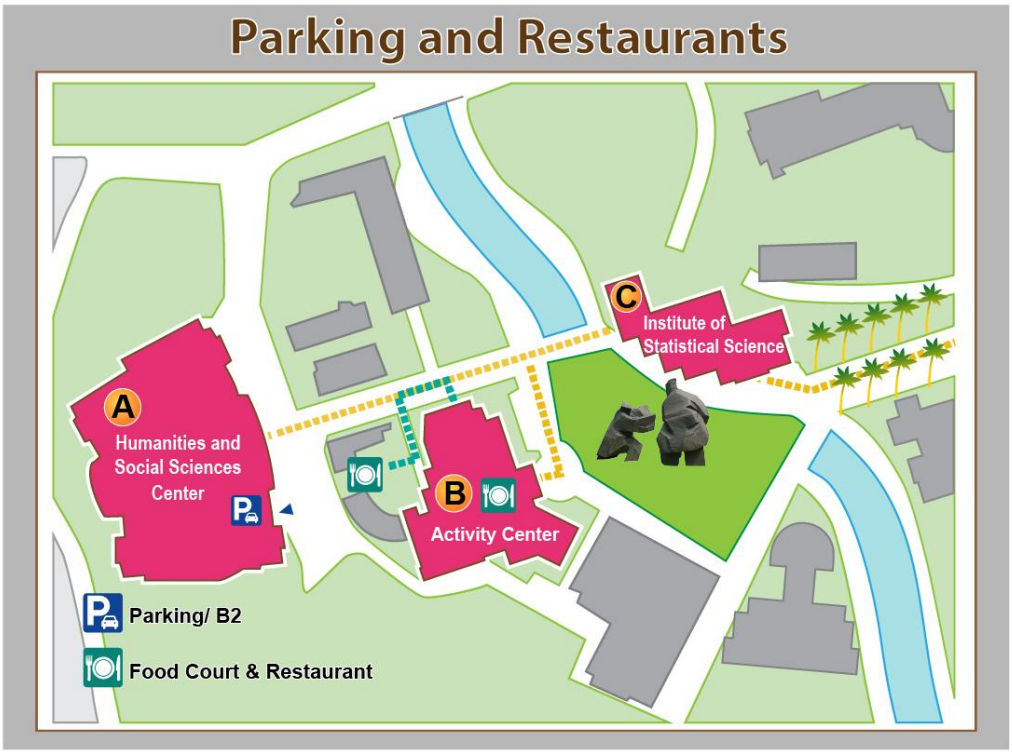

中央研究院北部院區地圖 Academia Sinica North Campus Map
國家生技研究園區
National Biotechnology Research Park
臺大院區
National Taiwan University
院區大門
Main Entrance
側門A
Entrance A
側門B
Entrance B
中研新村站
Zhongyan New Village Stop
中研院站
Academia Sinica Stop
胡適公園站
Hu Shih Park Stop
往國家生技研究園區
To National Biotechnology Research Park
生態池
Eco Pond
蔡元培紀念館
Tsai Yuan-Pei Memorial Hall
新温室大樓
Greenhouse Building
植物分子育種溫室
Plant Molecular Breeding Greenhouse
生物化學研究所
Institute of Biological Chemistry
分子生物研究所
Institute of Molecular Biology
生物醫學科學研究所
Institute of Biomedical Sciences
生態時代館
Eco Pavilion
生命科學圖書館
Life Science Library
基因體研究中心
Genomics Research Center
農業生物科技研究中心
Agricultural Biotechnology Research Center
植物暨微生物學研究所
Institute of Plant and Microbial Biology
白樓
White Building
植物標本館
Biodiversity Research Museum, Herbarium
黃樓
Yellow Tile Building
院本部行政大樓
Central Administrative Office
資訊科技創新研究中心
Research Center for Information Technology Innovation
中國文哲研究所
Institute of Chinese Literature and Philosophy
統計科學研究所
Institute of Statistical Science
人文社會科學研究中心
Research Center for Humanities and Social Sciences
化學研究所
Institute of Chemistry
物理研究所
Institute of Physics
資訊科學研究所
Institute of Information Science
近史所檔案館
Institute of Modern History Archives Buildings
胡適紀念館
Hu Shih Memorial Hall
嶺南美術館
Lingnan Fine Arts Museum
近代史研究所
Institute of Modern History
傅斯年圖書館
Fu Ssu-Nien Library
臺灣考古館
Taiwan Archaeology Building
經濟研究所
Institute of Economics
胡適公園
Hu Shih Park
歐美研究所
Institute of European and American Studies
歷史文物陳列館
Museum of the Institute of History and Philology
歷史語言研究所
Institute of History and Philology
數位文化中心
Academia Sinica Center for Digital Cultures
民族學研究所
Institute of Ethnology
民族學研究所博物館
Museum of the Institute of Ethnology
社會學研究所
Institute of Sociology
臺灣史研究所
Institute of Taiwan History
政治學研究所
Institute of Political Science
語言學研究所
Institute of Linguistics
法律學研究所
Institutum Iurisprudentiae
人文社會科學聯合圖書館
Joint Library of Humanities and Social Sciences
人文社會科學館
Humanities and Social Sciences Building(HSSB)
學術活動中心
Academic Activities Center
綜合體育館
Gymnasium
地球科學研究所
Institute of Earth Sciences
森林生態研究園區
Ecological Research Park
環境變遷研究大樓
Environmental Changes Research Building
環境變遷研究中心
Research Center for Environmental Changes
永續科學中心
Center for Sustainability Science
跨領域科技研究大樓
Interdisciplinary Research Building for Science and Technology
應用科學研究中心
Research Center for Applied Sciences
生物多樣性研究中心
Biodiversity Research Center
細胞與個體生物學研究所
Institute of Cellular and Organismic Biology
動物標本館
Biodiversity Research Museum, Zoological Collection
國際研究生大樓
TIGP Building
Counseling Center
中央研究院附設臺北市私立幼兒園
The Kindergarten of Academia Sinica
生醫轉譯研究中心
(國家生技研究園區C棟)
Biomedical Translational Research Center (National Biotechnology Research Park Building C)
環境教育中心
(國家生技研究園區F棟1樓)
Environmental Education Center (National Biotechnology Research Park Building F)
中研公園
Academia Park
天文數學館
Astronomy-Mathematics Bldg.
數學研究所
Institute of Mathematics
天文與天文物理研究所
Institute of Astronomy and Astrophysics
醉月湖
Drunken Moon Lake
原子與分子科學研究所
Institute of Atomic and Molecular Sciences
Google Map
中央研究院
ACADEMIA SINICA

Maps of Venues

Academia Sinica

https://maps.app.goo.gl/epNeaakWYn2vqKQS9

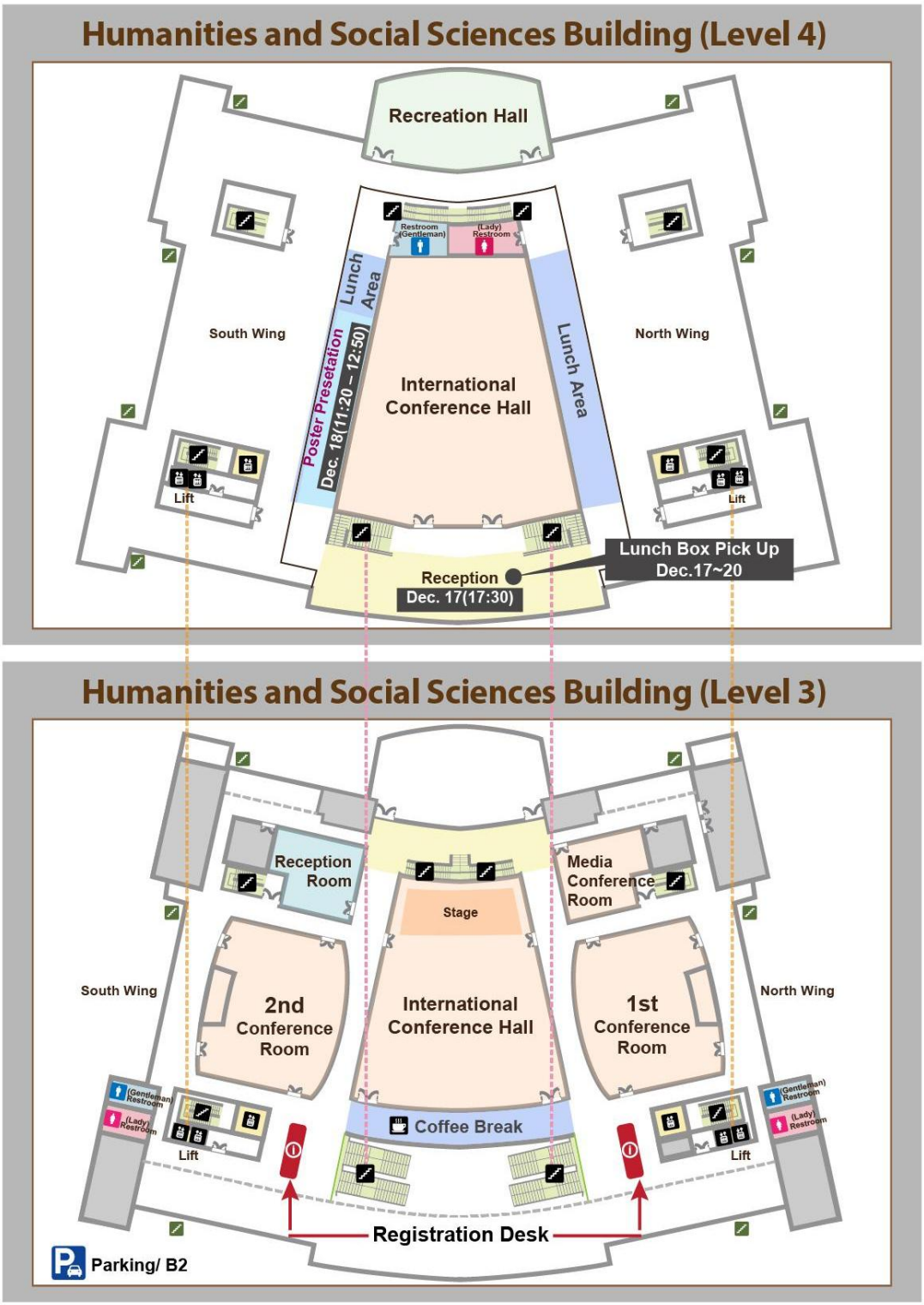




Floor Plans

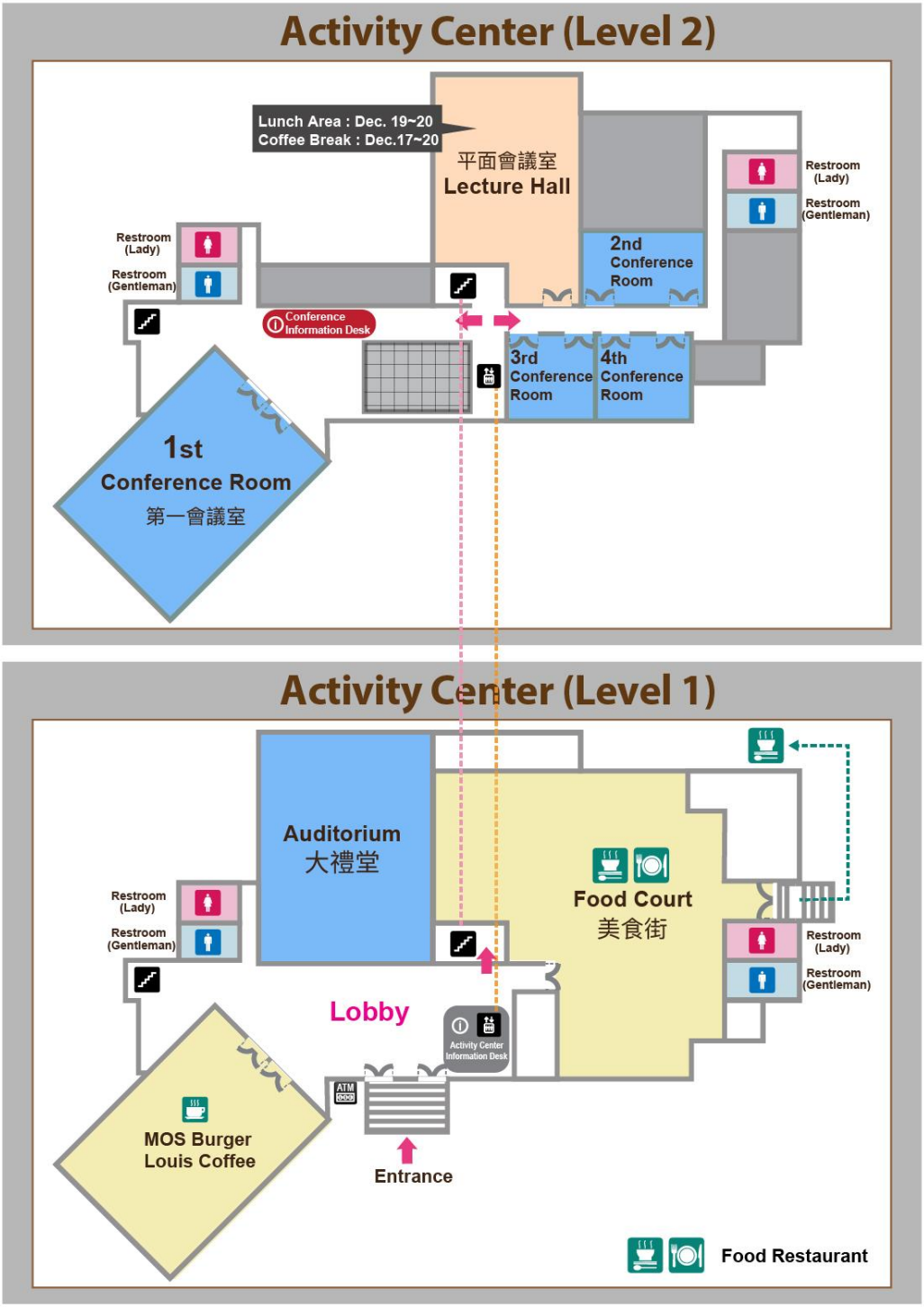
Humanities and Social Sciences Building (HSSB)

https://maps.app.goo.gl/G8dADnZzJnKsYi4n7

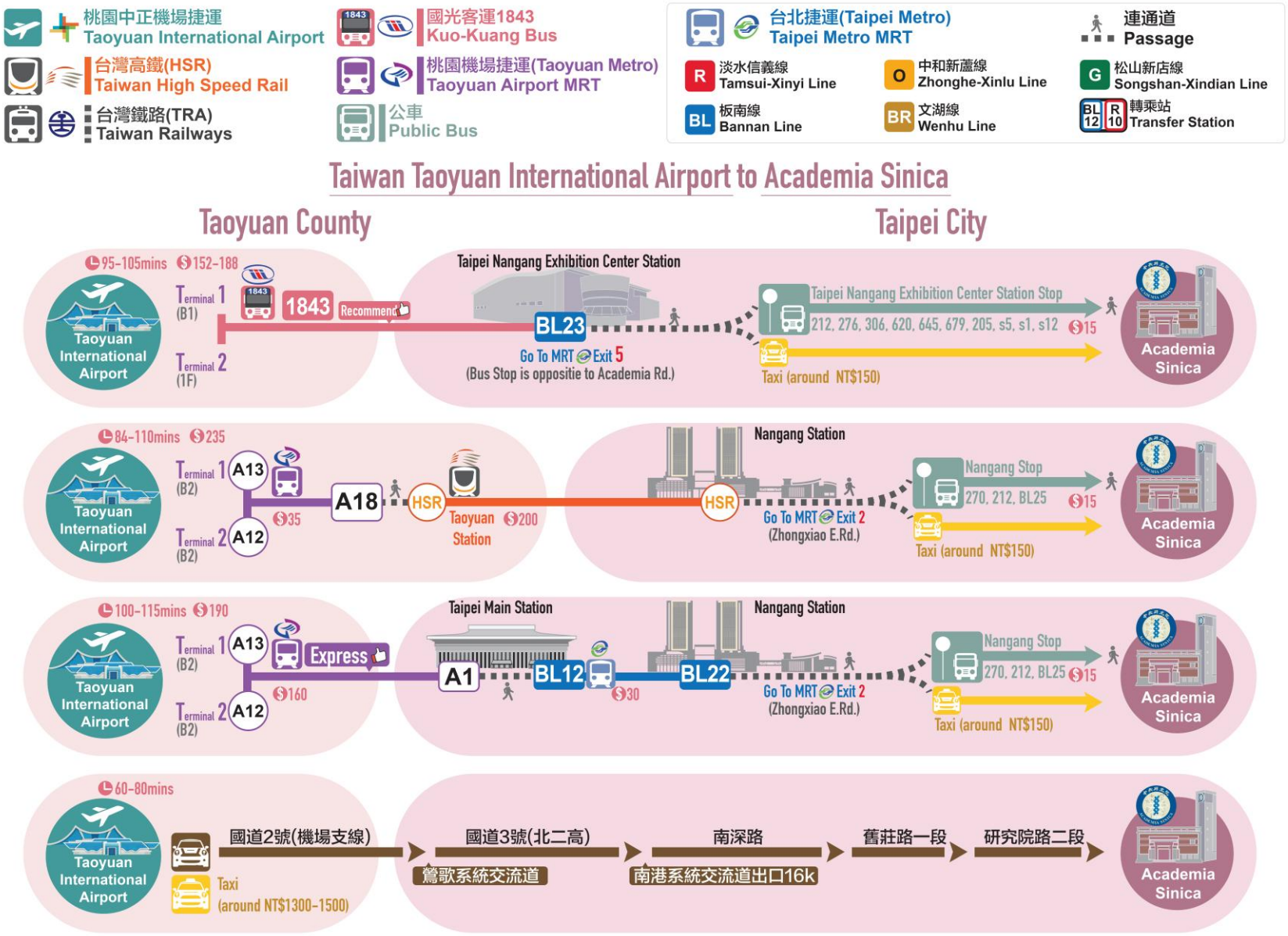


Activity Center (AC)

https://maps.app.goo.gl/xWKku4u2A6kd4rCD8



Transportation to Academia Sinica



A. By bus and taxi

1. Take the Kuo-Kuang Bus (Route 1843) to the Taipei Nangang Exhibition Center.
2. Go to Taipei Metro MRT Bannan Line (BL23) Exit 5. The bus stop is across Academia Road from Exit 5.
3. Take a bus or taxi to Academia Sinica. Please refer to the map for available bus routes.

B. By Taoyuan Airport MRT, Taiwan High Speed Rail and bus or taxi

1. Take the Taoyuan Airport MRT to Taoyuan High Speed Rail Station.
2. Take the Taiwan High Speed Rail to Nangang Station.
3. Go to Taipei Metro MRT Bannan Line (BL22) Exit 2. The bus stop is on Zhongxiao E. Rd.
4. Take a bus or taxi to Academia Sinica. Please refer to the map for available bus routes.

C. By Taoyuan Airport MRT, Taipei Metro MRT, and bus or taxi

1. Take the Taoyuan Airport MRT to Taipei Main Station.
2. Take the Taipei Metro MRT Bannan Line to Nangang station (BL22)
3. Go to Taipei Metro MRT Bannan Line (BL22) Exit 2. The bus stop is on Zhongxiao E. Rd.
4. Take a bus or taxi to Academia Sinica. Please refer to the map for available bus routes.

D. By car or taxi

1. Take National Freeway No. 2 (Airport Spur Route) and transfer via the Yingge System Interchange to National Freeway No. 3 (Formosa Freeway).
2. Continue via the Nangang System Interchange (Exit 16K) to Nanshen Road.
3. Follow Section 1 of Jiuzhuang Road, then Section 2 of Yanjiuyuan Road to reach Academia Sinica.

Accommodation

H1：The Place Taipei 南港老爺行旅

Telephone: +886-2-7750-0588
Fax: +886-2-2788-8582
Email: hrng.service@ng.hotelroyal.com.tw
Website: https://www.hotelroyal.com.tw/en-us/nangang
Location: Google Maps

H2：Taipei Forward Hotel Nangang Branch 馥華商旅南港館

Telephone: +886-2-2785-2655
Fax: +886-2-6615-6799
Email: forward@gmail.com
Website: https://fwhotelng.tw/
Location: Google Maps

H3：Green World NanGang 洛碁大飯店 南港館

Telephone: +886-2-2789-3009
Fax: +886-2-2789-3008
Email: nangang@gwh.global
Website: https://nangang.greenworldhotels.com/
Location: Google Maps

H4：Courtyard by Marriott Taipei 台北六福萬怡酒店

Telephone: +886-2-2171-6565
Fax: +886-2-2654-6565
Email: service@courtyardtaipei.com
Website: https://www.courtyardtaipei.com.tw/zh-TW/
Location: Google Maps

Accommodation Location

(Shuttle service from the hotels to the conference venue will be provided every morning.
Please check the official website or contact Conference Assistant for details.)

Social Events

Complimentary Half-Day Local Tour

Cultural Walk & Night-Market Feast
Date: Fri., Dec. 19, 2025
Duration: from 15:00 to 19:30
14:45 – 15:00 Meet at Academia Sinica · Departure
15:30 – 17:30 Songshan Cultural & Creative Park (Google Maps)
17:50 – 19:30 Raohe Night Market (Google Maps)
19:30 – 20:00 Return to Academia Sinica

It is our pleasure to invite you to visit Songshan Cultural and Creative Park and Raohe Night Market.

Songshan Cultural and Creative Park was originally the "Taiwan Governor-General's Tobacco Factory" during the Japanese colonial period. After restoration and revitalization, it has been transformed into a cultural park that blends historical architecture with creative industries. The park preserves red-brick factory buildings, boiler rooms, and chimneys from its industrial past, while integrating exhibition spaces, creative shops, cafés, and design brands. Today, it serves as an important venue for art exhibitions, creative markets, and design events—preserving Taiwan's industrial heritage while fostering a vibrant hub for young creative talents to showcase their work and exchange ideas.

Raohe Street Night Market is one of the city's most famous traditional night markets. Stretching about 600 meters long, it is well known for its wide variety of local delicacies such as pepper buns, oyster vermicelli, herbal pork rib soup, and "ice-hot" sweet rice balls, attracting both locals and tourists. In addition to food, the market also features stalls selling clothes, daily goods, and fun games, creating a lively and colorful atmosphere. At the entrance stands the historic Ciyou Temple, adding a rich touch of local culture to the night market experience.

Note: Attendees are required to wear their conference badge when boarding the bus and throughout the tour.

Reception and Banquet

Reception

The Buffet – 歐式自助式晚宴

Date: 2025/December/17 (Wednesday) 17:30
Location: HSSB Recreation Hall (4F), Academia Sinica
Map: Google Maps

Shuttlebus

From 18:30, depart from HSSB

- MRT route (every 10~15 mins): to MRT Taipei Nangang Exhibition Center (BL23)
- Hotel route (every 20 mins): H1 → H2 → H4 → H3

Join us for a buffet reception at Academia Sinica. A variety of dishes will be available, including appetizers, main courses, specialty snacks, and beverages. Feel free to help yourself and network with participants in this welcoming space.

Banquet

The Garden Taipei – 徐州路 2 號庭園會館

Date: 2025/December/18 (Thursday) 18:00
Location: No.2, Xuzhou Road, Zhongzheng District, Taipei City
Map: Google Maps

Shuttlebus

Outbound: depart from HSSB & AC at 16:50

Inbound: depart from **The Garden Taipei**

- Route 1: to AC, Academia Sinica
- Route 2: H2 → H1
- Route 3: H4 → H3

Join us for the conference banquet at The Garden Taipei, a classic restaurant in the heart of Taipei City. Enjoy an elegant Chinese-style round table dining experience, symbolizing unity and harmony, while being entertained by special performances. This banquet also serves as the annual gathering of the Institute of Statistical Science, Academia Sinica. All participants are warmly welcomed!

Conference Banquet Performance Program

3PEOPLEMUSIC

An Exploration of Tradition and Modernity

Founded in 2013, 3peoplemusic is a Taiwan-based trio blending classical roots with contemporary creativity. Featuring guzheng, zhongruan, and bamboo flutes, they craft a unique sound that bridges traditional Chinese music and global genres.

Their album, Change, won Best Crossover Music Album at the 33rd Golden Melody Awards for Traditional Arts and Music, and they earned a Bronze at the 2022 Global Music Awards. In 2025, 3peoplemusic received the Best Live Music Video award at the Europe Music Video Awards.

With performances across Asia, Europe and North America, they were invited to the 2024 Paris Cultural Olympiad, and in 2025, they became the first officially selected team to represent Taiwan at WOMEX, the world's largest world music expo in Finland. They continue to share Taiwan's dynamic new voice with audiences across the globe.

The Artists

KUO Jing-Mu, Leader and Guzheng (Chinese Zither)

JEN Chung, Musical Director and Dizi/Xiao (Bamboo Flutes)

PAN I-Tung, Music Producer and Ruan (Chinese Lute)

CHEN Yi-Ting, Administrator

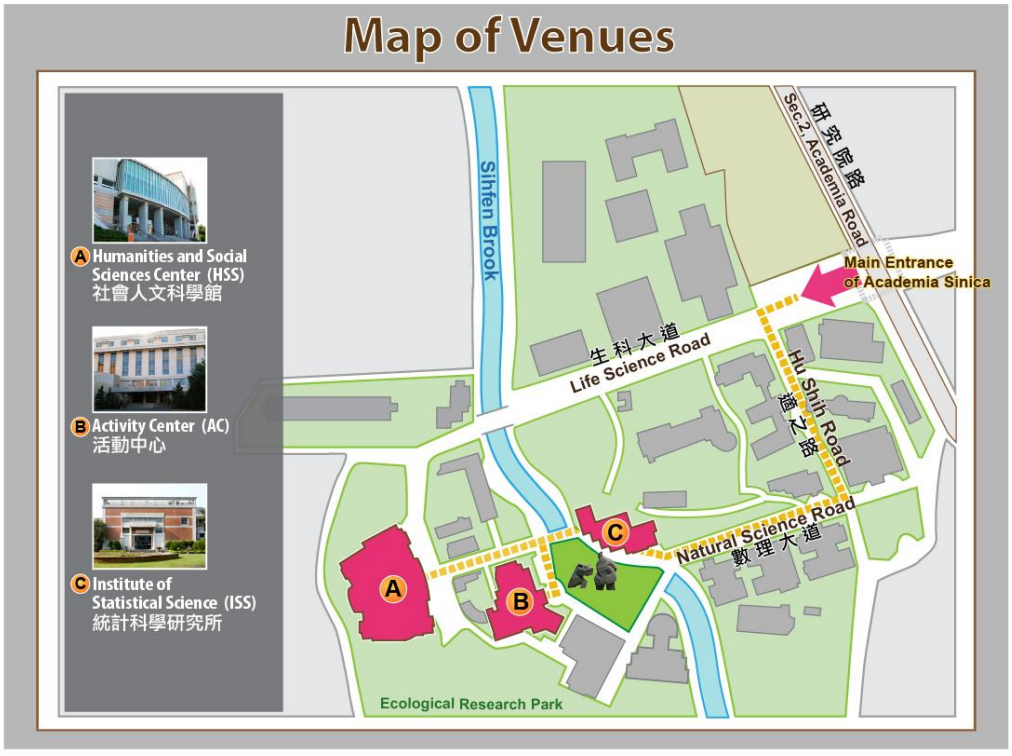
Schedule

Location ID

HSSB: https://maps.app.goo.gl/qK1SrSrPRb2jwMy1A
AC: https://maps.app.goo.gl/xWKku4u2A6kd4rCD8

S1: HSSB-3F-International Conference Hall（人文大樓三樓國際會議廳）
S2: HSSB-3F -1st Conference Room（人文大樓三樓第一會議室）
S3: HSSB-3F-2nd Conference Room（人文大樓三樓第二會議室）
S4: HSSB-3F-Media Conference Room（人文大樓三樓遠距會議室）
S5: HSSB-4F-Recreation Hall（人文大樓四樓交誼廳）

A1: AC-1F-Auditorium（活動中心一樓大禮堂）
A2: AC-2F-1st Conference Room（活動中心二樓第一會議室）
A3: AC-2F-2nd Conference Room（活動中心二樓第二會議室）
A4: AC-2F-3rd Conference Room（活動中心二樓第三會議室）
A5: AC-2F-4th Conference Room（活動中心二樓第四會議室）



Dec. 17 (Wednesday)

Click to access Speakers and Abstracts

08:00 – 09:00	Registration (HSSB)									
09:00 – 09:50	Welcome and Opening Ceremony (S1) AS Vice President **Mei-Yin Chou** + Group Photo + ISSAS Director **Hsin-Chou Yang** + ICSA President **Hongyu Zhao** + Conference Co-Chair **Ming-Chung Chang**									
09:50 – 10:50	Keynote Speech (S1) Chair: **Hongyu Zhao** Speaker: **Danyu Lin**									
10:50 – 11:10	Coffee Break									
11:10 – 12:10	Keynote Speech (S1) Chair: **Ming-Chung Chang** Speaker: **Jianqing Fan**									
12:10 – 13:10	Lunch Break (HSSB)									
Location	**S1**	**S2**	**S3**	**S4**	**S5**	**A1**	**A2**	**A3**	**A4**	**A5**
Organizer	Jane-Ling Wang		Huixia Judy Wang	Hongquan Xu	Zhengyuan Zhu	Mei-Ling Ting Lee	Andrew J. Holbrook	Yong Zang	Lei Liu Ying Ding	Hua Tang Nancy R. Zhang
13:10 – 14:50	17a1 Complex Analysis for Emerging Data		17a3 Innovations in Causal Mediation, Privacy, and AI-Driven Statistical Analysis	17a4 Experimental Design	17a5 Spatial Statistics	17a6 Big Data and AI	17a7 Quantum Computing in Statistics	17a8 Innovative Bayesian Methods for Adaptive Clinical Trials	17a9 Challenges in Survival Analysis and Clinical Trials	17a10 Genomic and Multi-Omic
14:50 – 15:10	Coffee Break									
Organizer	Henry Horng-Shing Lu	Hsin-Chou Yang	Hao Zhang	Lei Liu Ying Ding	Mei-Ling Ting Lee	Peihua Qiu	Takeshi Emura	Takeshi Emura Shuhei Ota	Cheng Zheng	(Tony) Jianguo Sun Guanyu Hu
15:10 – 16:50	17b1 Memorial Session for Prof. Shaw-Hwa Lo	17b2 Statistics and AI in Omics	17b3 Spatial Statistics and Machine Learning	17b4 Novel Methods for Event Prediction and Subgroup Identification	17b5 Recent Advances in Lifetime Data Analysis	17b6 Recent Research in Statistical Process Control, Part I	17b7 Survival Analysis with Frailty and Copulas	17b8 Recent Advances in Reliability Analysis	17b9 Advanced Methods for Causal Inference	17b10 Recent Advances in Bayesian Methods
16:50 –	Reception									

Dec. 18 (Thursday)

Click to access Speakers and Abstracts

08:00 – 09:00	Registration (HSSB)									
09:00 – 10:00	Keynote Speech (S1) Chair: **Hsin-Cheng Huang** Speaker: **Genevera I. Allen**									
10:00 – 10:20	Coffee Break									
10:20 – 11:20	Keynote Speech (S1) Chair: **Xinping Cui** Speaker: **Amy Xia**									
11:20 – 12:50	Poster Presentation (HSSB, 4F)									
	Lunch Break (HSSB)									
Location	**S1**	**S2**	**S3**	**S4**	**S5**	**A1**	**A2**	**A3**	**A4**	**A5**
Organizer	Henry Horng-Shing Lu	Ching-Kang Ing	Ying Hung		Peter Song	C. Jason Liang	George Michailidis	I-Ping Tu	Shujie Ma	Hongyuan Cao
12:50 – 14:30	18a1 Recent Developments for Data Science	18a2 Statistical Innovations for High-Dimensional and Time-Dependent Data	18a3 Analysis of Complex Data	18a4	18a5 New Data Analytics for Evaluating Complex Associations	18a6 Causal Inference for Survival Data	18a7 Recent Advances in Time Series Analysis	18a8 3D Protein Structure Informatics	18a9 New Developments in High-Dimensional Matrix and Network Analysis	18a10 Recent Methods Development in High-Dimensional Omics Studies
14:30 – 14:50	Coffee Break									
Organizer	Hsin-Cheng Huang	Lei Liu Ying Ding	Henry Horng-Shing Lu	Weng Kee Wong López-Fidalgo Jesús	Yi Li	Yichuan Zhao	Xinping Cui	Masato Abe	Takeshi Emura	Haiyan Huang Feng Liang
14:50 – 16:30	18b1 Spatial and Environmental Statistics	18b2 New Insights into Inference and Causality	18b3 Recent Developments for Biomedical Statistics	18b4 Recent Advances in Optimal Experimental Designs	18b5 High-Dimensional Models with Applications in Biomedical Sciences	18b6 Modern Machine Learning in the Big Data Era	18b7 Classical Meets Cutting-Edge: Regression, Mixtures, and Joint Models in Biomedical Research	18b8 Exploring Phenomena Through Mathematical Modeling	18b9 Copula and Dependence Modeling	18b10 Modern Bayesian Tools for Modeling and Inference
16:30 –	Banquet (The Garden Taipei)									

Dec. 19 (Friday)

Click to access Speakers and Abstracts

08:00 – 09:00	Registration (HSSB)									
09:00 – 10:00	ICSA Pao-Lu Hsu Award (S1) Chair: **Ying Zhang** Awardee/Speaker: **Hui Zou**									
10:00 – 10:20	Coffee Break									
Location	**S1**	**S2**	**S3**	**S4**	**S5**	**A1**	**A2**	**A3**	**A4**	**A5**
Organizer	Huixia Judy Wang	Yingying Fan	Xiaowu Dai	Cheng Zheng	Shujie Ma	Lei Liu Ying Ding	Xiaotong Shen	Guei-Feng (Cindy) Tsai	Jane-Ling Wang	Yuanjia Wang
10:20 – 12:00	19a1 Statistica Sinica Special Invited Papers	19a2 Frontiers in Statistical Inference: Dependence and Data Privacy	19a3 Advances in Causality, Reinforcement Learning, and Business Analytics	19a4 Advanced Methods for Novel Biomedical Data Types	19a5 Recent Advancements in Network and Correlated Data Analysis	19a6 Deep Learning and Artificial Intelligence	19a7 Recent Developments on Biostatistics and Ai	19a8 Recent Advances in Clinical Trials	19a9 Recent Advance in Statistics and AI	19a10 Modern Bayesian and Machine Learning Methods for Precision Medicine and Digital Health
12:00 – 12:50	Lunch Break (HSSB + AC)									
Organizer	Jane-Ling Wang	Juergen Symanzik	Yichuan Zhao	Nicolas Brunel	Xiulin Xie Peihua Qiu	Hongyuan Cao	Chun-Shu Chen/ Feng-Chang Lin	Tingting Zhang	Peng Ding	Ying Lu
12:50 – 14:30	19b1 What is Obscure about Random Objects?	19b2 Data Visualization	19b3 Recent Advances in Nonparametric Methods and Their Applications	19b4 Extending Canonical Conformal Predictions to Meet the Practitioners' Needs	19b5 Recent Research in Statistical Process Control, Part II	19b6 Recent Developments in AI	19b7 Recent Development of Statistical Methods in Case-Cohort Study Design and Dependent Sampling	19b8 Recent Research Developments in Neuroimaging Data Analysis	19b9 Causal Inference: Episode I	19b10 Recent Developments in Survival Analysis and Clinical Trial Methodology
14:30 –	Tour (see Local Tour)									

Dec. 20 (Saturday)

Click to access Speakers and Abstracts

08:00 – 8:40	Registration (HSSB)									
Location	S1	S2	S3	S4	S5	A1	A2	A3	A4	A5
Organizer	Hua Tang Nancy R. Zhang	John Stufken	Yen-Tsung Huang	Fangrong Yan	Emma Jingfei Zhang	Jin-Ting Zhang	Naitee Ting	Shuangge Ma Hao Mei	Yuchao Jiang	Yuxin Chen
08:40 – 10:20	20a1 Machine Learning / AI	20a2 Recent Advancements in Design of Experiments	20a3 Causal Inference	20a4 Bayesian Adaptive Designs for Oncology Dose Optimization	20a5 Recent Advances in Network and Tensor Data Analysis	20a6 Advances in Statistical Modelling and Inference for High-dimensional and Functional Data	20a7 Recent Advances in Clinical Trials	20a8 New Advances in Biostatistics and Bioinformatics	20a9 Innovations in Statistical Methodology for Complex Data: Spatial Omics, Stochastic Optimization, Network Integration, and Microscopy Image Analysis	20a10 Statistical and Algorithmic Foundations of Diffusion Models
10:20 – 10:40	Coffee Break									
Organizer	Ting-Li Chen	John Stufken	George Michailidis		Xingqiu Zhao	Jian Huang	I-Ping Tu	Guei-Feng (Cindy) Tsai		
10:40 – 12:20	20b1 Modern Functional Data Methods with Applications to Environmental Studies	20b2 Subsampling	20b3 Recent Advances in Time Series Analysis	20b4	20b5 Recent Developments in Survival Analysis and Deep Learning	20b6 Recent Advancements in Deep Learning and Graphical Models, and Model Selection	20b7 Applied Probability	20b8 Regulatory Advances in Clinical Trials	20b9	20b10
12:20 – 13:20	Lunch Break (HSSB+AC)									

Location	S1	S2	S3	S4	S5	A1	A2	A3	A4	A5
Organizer	Jane-Ling Wang	Su-Yun Huang	Henghsiu Tsai Mike K.P. So	Jian Huang	Haoda Fu	Xiaowu Dai	Haiyan Huang / Yuguo Chen	Zeny Feng	Maiying Kong	Nancy R. Zhang Fan Li
13:20 – 15:00	20c1 New Fronts on Machine Learning	20c2 Modern Advances in Learning, Robust Modeling, and Structure for High-Dimensional Data	20c3 Advances in Financial Econometrics and Network Modeling	20c4 Recent Advances in Data Science	20c5 Lead Science and Clinical Research – Career Panel Discussion	20c6 Flexible Inference: Nonparametrics, Causality, and Large Language Models	20c7 Recent Advances in Network Data Analysis	20c8 Complex Data Analysis in Environmental and Health Studies	20c9 Statistical Analyses Methods for Integrating Multi-Omics Data with Application to Personalized Medication	20c10 Causal Inference for Complex Designs
15:00 – 15:20	Coffee Break									
Organizer	Jane-Ling Wang	Masato Abe	Peter Song		Biao Cai Emma Jingfei Zhang	Yi Li	Boris Choy	Xiaofeng Zhu	Yuanjia Wang	
15:20 – 17:00	20d1 Causal Inference: Episode II	20d2 Frontiers in Knowledge Creation and Discovery	20d3 Advances in Stratified and Error-Prone Data Analysis	20d4	20d5 Recent Advances in Statistical Learning for Complex Data Structures	20d6 New Advances in Machine Learning and AI	20d7 Innovations in Machine Learning for Financial Data Analysis	20d8 Understanding the Biological Heterogeneity of Complex Traits through Omics Data	20d9 Innovations in Survival Analysis and Clinical Trials for Biomedical Research	20d10

Plenary Speakers

Danyu Lin

Department of Biostatistics, University of North Carolina at Chapel Hill, USA

Keynote Speech (I) 2025/December/17 (Wednesday) 09:50 – 10:50

Danyu Lin, Ph.D., is the Dennis Gillings Distinguished Professor of Biostatistics at the University of North Carolina at Chapel Hill. Dr. Lin has published 300 papers, with >50,000 citations and an h-index of 100. The statistical methods he developed have been used in thousands of scientific studies. His publications on COVID-19 vaccines and treatments (5 in New England Journal of Medicine, 4 in JAMA, and 2 in Lancet —all as corresponding author) have been viewed >1 million times, cited by the CDC, FDA, and WHO, and reported by The New York Times, The Washington Post, ABC News, and NBC News. Dr. Lin is an elected fellow of American Statistical Association and Institute of Mathematical Statistics and previously received Mortimer Spiegelman Award, George W. Snedecor Award, and ICSA Distinguished Achievement Award.

Evaluating the Effectiveness of COVID-19 Vaccines Over Time

Danyu Lin

Dennis Gillings Distinguished Professor, The University of North Carolina at Chapel Hill

ABSTRACT

Approximately 800 million COVID-19 cases and 7 million COVID-19 deaths have been reported to the World Health Organization thus far. Vaccination is a major tool to combat the COVID-19 pandemic, but its effectiveness wanes over time and tends to be lower against new SARS-CoV-2 variants. The knowledge about the waning effects of vaccination can guide boosting strategies. In a series of papers published in The New England Journal of Medicine and JAMA, we reported several large cohort studies using COVID-19 case surveillance and vaccination data from the states of North Carolina and Nebraska. We developed a novel statistical framework to evaluate the time-varying effects of the five generations of COVID-19 vaccines produced in the United States on infections with different SARS-CoV-2 variants and on severe outcomes (hospitalization and death). Our findings have been used by the World Health Organization and the U.S. Centers for Disease Control and Prevention and Food and Drug Administration and reported by The New York Times, The Washington Post, ABC News, and NBC News.

Keywords: B-spline; Cox model; Time-varying coefficients; Vaccination policy; Waning vaccine efficacy

Jianqing Fan

Department of Operations Research and Financial Engineering, Princeton University, USA

Keynote Speech (II) 2025/December/17 (Wednesday) 11:10 – 12:10

Dr. Jianqing Fan is the Frederick L. Moore ’18 Professor of Finance, Professor of Statistics, and Professor of Operations Research and Financial Engineering at Princeton University, where he chaired the department from 2012 to 2015. He is a statistician, financial econometrician, and data scientist who has made influential contributions to statistical machine learning, data science, and their applications in finance, economics, computational biology, and biostatistics. He has particular expertise in high-dimensional statistics, spectral methods, neural networks, reinforcement learning, and survival analysis, among others. Dr. Fan has coauthored four highly regarded books—Local Polynomial Modeling and Its Applications (1996), Nonlinear Time Series: Parametric and Nonparametric Methods (2003), Elements of Financial Econometrics (2015), and Statistical Foundations of Data Science (2020)—and authored or coauthored over 300 peer-reviewed articles in leading journals across finance, economics, statistical machine learning, and other aspects of theoretical and methodological statistics. He is currently a joint editor of the Journal of American Statistical Association (2023 -2026) and has served as co-editor(-in-chief) for many top journals, including the Annals of Statistics (2004 - 2006), Probability Theory and Related Fields (2003 - 2005), Econometrical Journal (2007 - 2012), Journal of Econometrics (2012–2018), Journal of Business and Economics Statistics (2018-2021), and others. He is a past president of both the Institute of Mathematical Statistics and the International Chinese Statistical Association. Dr. Fan’s outstanding achievements have been recognized by numerous honors, including the COPSS President Award (2000), Guggenheim Fellowship (2009), Pao-Lu Hsu Prize (2013), Guy Medal in Silver (2014), Noether Distinguished Scholar Award (2018), Le Cam Award and Lectures (2021), Wald Award and Lectures (2025), among others. He was elected to Academician from Academia Sinica in 2012 and Royal Flemish Academy of Belgium in 2023.

Factor Informed Double Deep Learning for Average Treatment Effect Estimation

Jianqing Fan, Soham Jana, Sanjeev Kulkarni, and Qishuo Yin

Department of Operations Research and Financial Engineering, Princeton University

ABSTRACT

We investigate the problem of estimating the average treatment effect (ATE) under a very general setup where the covariates can be high-dimensional, highly correlated, and can have sparse nonlinear effects on the propensity and outcome models. We present the use of a Double Deep Learning strategy for estimation, which involves combining recently developed factor-augmented deep learning-based estimators, FAST-NN, for both the response functions and propensity scores to achieve our goal. By using FAST-NN, our method can select variables that contribute to propensity and outcome models in a completely nonparametric and algorithmic manner and adaptively learn low-dimensional function structures through neural networks. Our proposed novel estimator, FIDDLE (Factor Informed Double Deep Learning Estimator), estimates ATE based on the framework of augmented inverse propensity weighting AIPW with the FAST-NN-based response and propensity estimates. FIDDLE consistently estimates ATE even under model misspecification, and is flexible to also allow for low-dimensional covariates. Our method achieves semiparametric efficiency under a very flexible family of propensity and outcome models. We present extensive numerical studies on synthetic and real datasets to support our theoretical guarantees and establish the advantages of our methods over other traditional choices, especially when the data dimension is large.

Keywords: Factor models, Deep learning, FAST-NN, AIPW, Average treatment effect

Genevera Allen

Department of Statistics, Columbia University, USA

Keynote Speech (III) 2025/December/18 (Thursday) 09:00 – 10:00

Genevera Allen is a Professor of Statistics at Columbia University and a member of the Center for Theoretical Neuroscience, the Zuckerman Institute for Mind, Brain, and Behavior, and the Irving Institute for Cancer Dynamics. Before joining Columbia, Dr. Allen spent fourteen years at Rice University in the Departments of Electrical and Computer Engineering, Statistics, and Computer Science, where she founded and directed Rice's data science education center, informally known as the Rice D2K Lab. Her research focuses on developing statistical machine learning tools to help people make reliable discoveries from data, with contributions in unsupervised learning, interpretable machine learning, data integration, graphical models, and high-dimensional statistics, often motivated by neuroscience and bioinformatics. She has received several honors, including the National Science Foundation Career Award, Rice University's Duncan Achievement Award for Outstanding Faculty, and recognition in "*Forbes* '30 under 30′: Science and Healthcare" (2014). Dr. Allen is also an elected fellow of the American Statistical Association, the Institute of Mathematical Statistics, and the International Statistics Institute. She serves as an Action Editor for high-impact journals such as *JMLR* and *JASA*. Dr. Allen earned her Ph.D. in Statistics from Stanford University under Prof. Robert Tibshirani and her bachelor's degree, also in Statistics, from Rice University.

Inference for Interpretable Machine Learning: Feature Importance and Beyond

Genevera I. Allen

Department of Statistics, Columbia University, USA

ABSTRACT

Machine Learning (ML) systems are being used to make critical societal, scientific, and business decisions. To promote trust, transparency, and accountability in these systems, many advocate making them interpretable or explainable. In response, there has been dramatic growth in techniques to provide human understandable interpretations of black-box techniques. Yet we ask: Can we trust these ML interpretations? How do we know if they are correct? Unlike for prediction tasks, it is difficult to directly test the veracity of ML interpretations. In this talk, we focus on interpreting predictive models to understand important features and important feature patterns. We first present motivating results from a large-scale empirical stability study illustrating that feature interpretations are generally unreliable and far less reliable than predictions. Motivated by these issues, we present a new statistical inference framework for quantifying the uncertainty in feature importance and higher-order feature patterns. Based upon the Leave-One-Covariate-Out (LOCO) framework, we develop a computational and inferential approach that does not require data splitting or model refitting by utilizing minipatch ensembles, or ensembles generated by double random subsampling of observations and features. Even though our framework uses the same data for training and inference, we prove the asymptotic validity of our confidence intervals for LOCO feature importance under mild assumptions. Finally, we extend our approach to detect and test feature interactions via the iLOCO metric. Our approach allows one to test whether a feature significantly contributes to any ML model's predictive ability in a completely distribution free manner, thus promoting trust in ML feature interpretations. We highlight our inference for interpretable ML approaches via real scientific case studies and a fun illustrative example. This is joint work with Lili Zheng, Luqin Gan, Camille Little, Tarek Zikry, and Mariah Loehr.

Keywords: Conformal Inference, Selective Inference, Feature Importance Inference, Feature Interaction, Interpretable Machine Learning, Ensemble Learning

Amy Xia

Vice President, Amgen Inc., USA

Keynote Speech (IV) 2025/December/18 (Thursday) 10:20 – 11:20

Amy Xia is Vice President, Center for Design and Analysis at Amgen (including Biostatics, Statistical Programming, Design & Innovation, Data Sciences, etc), providing leadership for strategic drug development and driving innovative program/study designs across Amgen's portfolios for evidence generation and decision-making. Amy has extensive experience in biopharmaceutical and medical device industries in the past two decades across all development phases and a variety of disease areas. Amy is a recognized industry expert in innovative Bayesian applications, adaptive designs, pediatric extrapolation, meta-analysis and safety evaluation in drug development, and pioneers work in design optimization, evidence synthesis and product/portfolio decision-making. She is currently a member of the ICH E20 Expert Working Group on Adaptive Clinical Trials and previously served as the Vice-Chair for the DIA Bayesian Scientific Working Group. Amy is an elected Fellow of American Statistical Association. She received her PhD in Biostatistics from the University of Minnesota.

Shaping the Future: The Expanding Role of Pharma Statisticians in the AI Era

Amy Xia

Amgen, USA

ABSTRACT

In today's rapidly evolving pharmaceutical landscape, statisticians have become catalysts for innovation, turning vast and complex data into discoveries that change lives. No longer confined to the back room of clinical trial analysis, statisticians now stand at the center of drug discovery and development—guiding decisions, shaping strategies, and driving breakthroughs across the entire value chain. With the rise of real-world data, genomics, digital health, and artificial intelligence, our profession is redefining what is possible: accelerating trials, unlocking hidden insights, and bringing safer, more effective treatments to patients faster. This transformation also challenges us to grow—expanding our skills, embracing collaboration, and leading with both scientific rigor and vision. As we look ahead, the future of pharmaceutical innovation will be written by those who can combine statistical thinking with bold new technologies. Statisticians are not just keeping pace with change—we are shaping the future of medicine.

Key Words: Adaptive designs; Bayesian methods; Complex Innovative Designs; Artificial Intelligence; Data science; Real-world evidence

Hui Zou

School of Statistics, University of Minnesota, USA

ICSA Pao-Lu Hsu Award 2025/December/19 (Friday) 09:00 – 10:00

Hui Zou received his Ph.D. in Statistics from Stanford University and is Dr. Lynn Y.S. Lin Distinguished professor at the University of Minnesota. His primary research interests include statistical learning, flexible statistical modeling, and statistical computing. He has received numerous honors, including the NSF career Award (2009), IMS Tweedie Award (2011), Scholar of the College (2015), Best Paper Award in Applied Mathematics at the 2019 ICCM, 2025 Founders of Statistics Prize, and the ICSA Pao-Lu Hsu Award (2025). He is an elected Fellow of the American Statistical Association, the Institute of Mathematical Statistics, and the American Association for the Advancement of Science. From 2014–2019, Hui has been identified as an ISI Highly Cited Researcher in mathematics, and from 2020–2024, he was ranked among the world's top 2% of scientists by Stanford University. He has also authored or co-authored influential book chapters on variable selection for the linear support vector machine (2007) and high-dimensional classification (2018), contributing broadly to the theory and applications of modern statistical learning.

High-Dimensional Clustering via a Latent Transformation Mixture Model

Hui Zou

School of Statistics, University of Minnesota

ABSTRACT

Cluster analysis is a fundamental task in machine learning. From the probabilistic modeling viewpoint, a finite mixture model is naturally suited for the distribution of data with multiple clusters, and hence model-based clustering (MBC) offers an effective solution. Despite its many successful applications, MBC also often underperforms due to its potential severe modeling bias. We aim to design a more robust off-the-shelf MBC for high-dimensional data by mitigating the model bias. To this end, we propose a novel CESME model by incorporating nonparametric latent transformations into the finite Gaussian mixture model (GMM). The inclusion of latent transformations significantly enhances the flexibility of the finite GMM without compromising interpretability. We derive a model fitting procedure for implementing the optimal clustering under the CESME model and analyze the clustering accuracy of the resulting algorithm. It is shown that the additional cost due to estimating nonparametric transformations is negligible compared with an ideal clustering algorithm with known transformations. On six benchmark single-cell RNA sequence datasets, CESME exhibits dominating advantages over existing methods in the literature.

Keywords: Model-based clustering, Nonparametric transformation, High-dimensional data

Daily Schedule

December 17 (Wednesday):

Keynote Speech (I) [09:50 – 10:50]:

Location: S1
Title: ***Evaluating the Effectiveness of COVID-19 Vaccines Over Time***
Speaker: **Danyu Lin** (University of North Carolina at Chapel Hill, USA)
Chair: **Hongyu Zhao** (Yale University, USA)

Keynote Speech (II) [11:10 – 12:10]:

Location: S1
Title: ***Factor Informed Double Deep Learning for Average Treatment Effect Estimation***
Speaker: **Jianqing Fan** (Princeton University, USA)
Chair: **Ming-Chung Chang** (Academia Sinica, Taiwan, ROC)

Parallel Sessions [Dec. 17, 13:10 – 14:50]:

17a1 - Complex Analysis for Emerging Data	
Location: S1	To Abstracts
Organizer: **Jane-Ling Wang** (University of California, Davis, USA)	
Chair: **Jane-Ling Wang** (University of California, Davis, USA)	
13:10-13:35	*Synthetic Data–Powered Statistical Inference with Generative AI* **Xihong Lin** (Harvard University, USA)
13:45-14:10	*Additive Frechet Regression for Random Objects* **Byeong U. Park** (Seoul National University, South Korea)
14:20-14:45	*Goodness-of-Fit and the Best Approximation: An Adversarial Approach* **Qiwei Yao** (London School of Economics, United Kingdom)

Parallel Sessions [Dec. 17, 13:10 – 14:50]

17a3 - Innovations in Causal Mediation, Privacy, and AI-Driven Statistical Analysis	
Location: S3	To Abstracts
Organizer: **Huixia Judy Wang** (Rice University, USA)	
Chair: **Huixia Judy Wang** (Rice University, USA)	
13:10-13:35	*Assessing Spatial Spillover Effects in Mediation Analysis with Areal Data* **Peter Song** (University of Michigan, USA)
13:45-14:10	*Differentially Private Inference for Longitudinal Linear Regression* **Marco Avella Medina** (Columbia University, USA)
14:20-14:45	*Learning from Shifted Data via a Semiparametric Selection Bias Model* **Jiayang Sun** (George Mason University, USA)

17a4 - Experimental Design	
Location: S4	To Abstracts
Organizer: **Hongquan Xu** (University of California, Los Angeles, USA)	
Chair: **Cheng-Yu Sun** (National Tsing Hua University, Taiwan, ROC)	
13:10-13:30	*Orthogonalized Moment Aberration for Multi-Stratum Factorial Designs* **Ming-Chung Chang** (Academia Sinica, Taiwan, ROC)
13:35-13:55	*Efficient and Robust Block Designs for Order-of-Addition Experiments* **Chang-Yun Lin** (National Chung Hsing University, Taiwan, ROC)
14:00-14:20	*Results on Large Strong Orthogonal Arrays of Strength Three* **Chenlu Shi** (New Jersey Institute of Technology, USA)
14:25-14:45	*A Stratified L_2-Discrepancy with Application to Space-Filling Designs* **Hongquan Xu** (University of California, Los Angeles, USA)

Parallel Sessions [Dec. 17, 13:10 – 14:50]

17a5 - Spatial Statistics	
Location: S5	To Abstracts
Organizer: **Zhengyuan Zhu** (Iowa State University, USA)	
Chair: **Junho Yang** (Academia Sinica, Taiwan, ROC)	
13:10-13:35	*Asymmetric Space–Time Covariance Functions via Hierarchical Mixtures* **Pulong Ma** (Iowa State University, USA)
13:45-14:10	*Graphical Modeling of Multivariate Inhomogeneous Spatial Point Processes* **Junho Yang** (Academia Sinica, Taiwan, ROC)
14:20-14:45	*Fast Variable Selection in Semiparametric Spatial Zero-inflated Models: Application to Extreme Climate Events* **Chun-Shu Chen** (National Central University, Taiwan, ROC)

17a6 - Big Data and AI	
Location: A1	To Abstracts
Organizer: **Mei-Ling Ting Lee** (University of Maryland, USA)	
Chair: **George Tseng** (University of Pittsburgh, USA)	
13:10-13:35	*Leveraging AI Techniques to Study the Risk of Accelerated Brain Aging and Dementia Using Large-Scale Biobank Data* **Tianzhou (Charles) Ma** (University of Maryland, USA)
13:45-14:10	*On MCMC Mixing for Predictive Inference under Unidentified Transformation Models* **Catherine Chunling Liu** (The Hong Kong Polytechnic University, Hong Kong)
14:20-14:45	*Microbiome Data Integration via Shared Dictionary Learning* **Shulei Wang** (University of Illinois Urbana-Champaign, USA)

Parallel Sessions [Dec. 17, 13:10 – 14:50]

17a7 - Quantum Computing in Statistics	
Location: A2	To Abstracts
Organizer: **Andrew J. Holbrook** (University of California, Los Angeles, USA)	
Chair: **Andrew J. Holbrook** (University of California, Los Angeles, USA)	
13:10-13:30	*Quantum Speedups for Multiproposal MCMC* **Chin-Yi Lin** (Foxconn Quantum Research Center, Taiwan, ROC)
13:35-13:55	*A Derivative-Free Approach for Parameter Inference in Hidden Quantum Markov Models* **Patricia Ning** (Texas A&M University, USA)
14:00-14:20	*Quantum Computations of Partial Differential Equations and Related Problems* **Shi Jin** (Shanghai Jiao Tong University, China)
14:25-14:45	*Adaptive Circuit Learning of Born Machine: Towards Realization of Amplitude Embedding and Quantum Data Loading* **Hao-Chung Cheng** (National Taiwan University, Taiwan, ROC)

17a8 - Innovative Bayesian Methods for Adaptive Clinical Trials	
Location: A3	To Abstracts
Organizer: **Yong Zang** (Indiana University, USA)	
Chair: **Ying Yuan** (University of Texas MD Anderson Cancer Center, USA)	
13:10-13:35	*Randomized Optimal Selection Design for Dose Optimization* **Ying Yuan** (The University of Texas M.D. Anderson Cancer Center, USA)
13:45-14:10	*Exploring Sensitive Biomarkers with Short-Term Response and Long-term Outcome using Bayesian Additive Regression Trees* **Satoshi Morita** (Kyoto University Graduate School of Medicine, Japan)
14:20-14:45	*BIT: A Bayesian Optimal Adaptive Clinical Trial Design for Integrated Therapies* **Yong Zang** (Indiana University, USA)

Parallel Sessions [Dec. 17, 13:10 – 14:50]

17a9 - Challenges in Survival Analysis and Clinical Trials	
Location: A4	To Abstracts
Organizer: **Lei Liu** (Washington University in St. Louis, USA) **Ying Ding** (University of Pittsburgh, USA)	
Chair: **Takeshi Emura** (Hiroshima University, Japan)	
13:10-13:35	*An Empirical Bayesian Method for Subgroup Identification in Personalized Medicine* **Jingwei Wu** (Temple University, USA)
13:45-14:10	*Dynamic and Concordance-Assisted Learning for Risk Stratification* **Jing Ning** (The University of Texas M.D. Anderson Cancer Center, USA)
14:20-14:45	*Model Based Multiple Imputation in Censored Quantile Regression* **Zhaozhi Fan** (Memorial University of Newfoundland, Canada)

17a10 - Genomic and Multi-Omic	
Location: A5	To Abstracts
Organizer: **Hua Tang** (Stanford University, USA) **Nancy R. Zhang** (University of Pennsylvania, USA)	
Chair: **Marc Coram** (Google Zurich, Switzerland)	
13:10-13:30	*Improving Single-Cell Perturbation Analyses through Efficiency Estimation* **Jingshu Wang** (University of Chicago, USA)
13:35-13:55	*Addressing Heterogeneous Sensitivity in Biomarker Screening with Application in NanoString nCounter Data* **Zhijin Jean Wu** (Brown University, USA)
14:00-14:20	*The Taiwan Precision Medicine Initiative: Building the Largest non-European Cohort for Precision Health* **Jer-Yuarn Wu** (Academia Sinica, Taiwan, ROC)
14:25-14:45	*A Multi-Tissue Map of Protein Regulation Reveals Shared and Context-Dependent Genetic Architectures* **Hua Tang** (Stanford University, USA)

Back to Sessions List

Parallel Sessions [Dec. 17, 15:10 – 16:50]:

17b1 - Memorial Session for Prof. Shaw-Hwa Lo	
Location: S1	
Organizer: **Henry Horng-Shing Lu** (Kaohsiung Medical University / National Yang Ming Chiao Tung University, Taiwan, ROC)	
Chair: **Ray-Bing Chen** (National Tsing Hua University, Taiwan, ROC)	
1	**Jane-Ling Wang** (University of California, Davis, USA)
2	**Mong-Na Lo Huang** (National Sun Yat-Sen University, Taiwan, ROC)
3	**Xiao-Li Meng** (Harvard University, USA)
4	**Xihong Lin** (Harvard University, USA)
5	**Zhezhen Jin** (Columbia University, USA)
6	**Kani Chen** (The Hong Kong University of Science and Technology, Hong Kong)
7	**Henry Horng-Shing Lu** (Kaohsiung Medical University / National Yang Ming Chiao Tung University, Taiwan, ROC)

17b2 - Statistics and AI in Omics	
Location: S2	To Abstracts
Organizer: **Hsin-Chou Yang** (Academia Sinica, Taiwan, ROC)	
Chair: **Hsin-Chou Yang** (Academia Sinica, Taiwan, ROC)	
15:10-15:35	*Bayesian Rhythmic Model for Jointly Detecting Circadian Biomarkers and Predicting Molecular Circadian Time in Human Post-Mortem Brain Transcriptome* **George Tseng** (University of Pittsburgh, USA)
15:45-16:10	*Adjusting Transcript Leakage in Spatial Transcriptomic Data* **Yingying Wei** (The Chinese University of Hong Kong, Hong Kong)
16:20-16:45	*Computational Intelligence from Omics to Medicine* **Ka-Chun Wong** (City University of Hong Kong, Hong Kong)

Back to Sessions List

Parallel Sessions [Dec. 17, 15:10 – 16:50]

17b3 - Spatial Statistics and Machine Learning	
Location: S3	To Abstracts
Organizer: **Hao Zhang** (Michigan State University, USA)	
Chair: **Whitney Huang** (Clemson University, USA)	
15:10-15:30	*Estimation and Selection in Survival Models for Individuals with Spatial Frailty* **Chae Young Lim** (Seoul National University, South Korea)
15:35-15:55	*A Spatio-Temporal Modeling Approach for Wind Speed Data from a Regional Climate Model* **Whitney Huang** (Clemson University, USA)
16:00-16:20	*A Comparative Study of Neural Network Adaptations for Spatial Data* **Bo-yu Chen** (Purdue University, USA)
16:25-16:45	*Spectral Radii of Kernel Matrices and Applications to Kernel Score Tests* **Hao Zhang** (Michigan State University, USA)

17b4 - Novel Methods for Event Prediction and Subgroup Identification	
Location: S4	To Abstracts
Organizer: **Lei Liu** (Washington University in St. Louis, USA) **Ying Ding** (University of Pittsburgh, USA)	
Chair: **Jing Ning** (The University of Texas M.D. Anderson Cancer Center, USA)	
15:10-15:35	*Dynamic and Individualized Prediction of Cardiovascular Events: The International Childhood Cardiovascular Cohort (i3C) Consortium* **Nanhua Zhang** (University of Cincinnati / Cincinnati Children's Hospital Medical Center, USA)
15:45-16:10	*OPERA: An Interpretable Algorithm for Patient Stratification based on Partially Ordered Risk Factors* **Menggang Yu** (University of Michigan, USA)
16:20-16:45	*α-Separability and Adjustable Combination of Amplitude and Phase Model for Functional Data* **Jimin Ding** (Washington University in St. Louis, USA)

Parallel Sessions [Dec. 17, 15:10 – 16:50]

17b5 - Recent Advances in Lifetime Data Analysis	
Location: S5	To Abstracts
Organizer: **Mei-Ling Ting Lee** (University of Maryland, USA)	
Chair: **Gang Li** (University of California, Los Angeles, USA)	
15:10-15:30	*Efficient Estimation for Recurrent Events under Informative Censoring Using Generalized Method of Moments* **Yu-Jen Cheng** (National Tsing Hua University, Taiwan, ROC)
15:35-15:55	*A Doubly Robust Instrumental Variable Approach for Estimating Average Treatment Effects in Time-to-Event Data with Unmeasured Confounding* **Chung-Chou Ho Chang** (University of Pittsburgh, USA)
16:00-16:20	*A Quantile Cure Model with Partially Functional Covariate Effects* **Chyong-Mei Chen** (National Yang Ming Chiao Tung University, Taiwan, ROC)
16:25-16:45	*Semiparametric Analysis of Multivariate Panel Count Data with Informative Observation Processes* **Xin He** (University of Maryland, College Park, USA)

17b6 - Recent Research in Statistical Process Control, Part I	
Location: A1	To Abstracts
Organizer: **Peihua Qiu** (University of Florida, USA)	
Chair: **Peihua Qiu** (University of Florida, USA)	
15:10-15:30	*Using Interpoint Distances to Develop a New Multivariate Control Chart Based on Change-Point Detection* **Claudio Borroni** (University of Milano – Bicocca, Italy) **Manuela Cazzaro** (University of Milano – Bicocca, Italy)
15:35-15:55	*Some Change-Point Design-Based Distribution-free Approaches for Monitoring High-Dimensional Data* **Amitava Mukherjee** (XLRI - Xavier School of Management, India)
16:00-16:20	*Large-Scale Decentralized Fault Diagnosis for Multi-Group Data with Auxiliary Information via Distributed Multiple Testing* **Zhihan Zhang** (East China Normal University, China)
16:25-16:45	*Control Chart for High-Dimensional Dynamic Process Monitoring* **Peihua Qiu** (University of Florida, USA)

Back to Sessions List

Parallel Sessions [Dec. 17, 15:10 – 16:50]

17b7 - Survival Analysis with Frailty and Copulas	
Location: A2	To Abstracts
Organizer: **Takeshi Emura** (Hiroshima University, Japan)	
Chair: **Takeshi Emura** (Hiroshima University, Japan)	
15:10-15:30	*Survival Models with a Cured Fraction: A Zero-Inflated Gamma Frailty–Copula Approach* **Masaki Hino** (The Graduate University for Advanced Studies, SOKENDAI, Japan)
15:35-15:55	*Mean Residual Life Based Illness-Death Model for Semicompeting Risks Data* **Liming Xiang** (Nanyang Technological University, Singapore)
16:00-16:20	*H-Likelihood Approach on the Joint Frailty Model for Clustered Bivariate Survival Data* **IL-Do Ha** (Pukyong National University, South Korea)
16:25-16:45	*Inferring Median Survival under Dependent Censoring* **Takeshi Emura** (Hiroshima University, Japan)

17b8 - Recent Advances in Reliability Analysis	
Location: A3	To Abstracts
Organizer: **Takeshi Emura** (Hiroshima University, Japan) **Shuhei Ota** (Kanagawa University, Japan)	
Chair: **Shuhei Ota** (Kanagawa University, Japan)	
15:10-15:30	*Degradation Models for Life Time Estimation of Serial and Parallel Connected Lithium-ion Battery Packs* **Shuen-Lin Jeng** (National Cheng Kung University, Taiwan, ROC)
15:35-15:55	*Optimal Designs for Gamma Degradation Tests* **Hung-Ping Tung** (National Yang Ming Chiao Tung University, Taiwan, ROC)
16:00-16:20	*Shrinkage Estimation for the Rate Parameter under the Exponential Distribution with Censored Survival Data* **Nanami Taketomi** (Nagasaki University, Japan)
16:25-16:45	*Reliability Analysis for Small Ball Bearings by Considering the Correlation of Their Lifetimes* **Shuhei Ota** (Kanagawa University, Japan)

Parallel Sessions [Dec. 17, 15:10 – 16:50]

17b9 - Advanced Methods for Causal Inference	
Location: A4	To Abstracts
Organizer: **Cheng Zheng** (University of Nebraska Medical Center, USA)	
Chair: **Hsueh-Han Huang** (Academia Sinica, Taiwan, ROC)	
15:10-15:35	*A Simple Nonparametric Least-Squares-Based Causal Inference for Heterogeneous Treatment Effects* **Ying Zhang** (University of Nebraska Medical Center, USA)
15:45-16:10	*Using Negative Controls to Adjust for Unmeasured Confounding in Continuous Exposure Settings* **Jie Kate Hu** (The Ohio State University, USA)
16:20-16:45	*Causal Mediation Analysis for Survival Outcome and Recurrent Event Mediators with Time-Varying Confounding* **Cheng Zheng** (University of Nebraska Medical Center, USA)

17b10 - Recent Advances in Bayesian Methods	
Location: A5	To Abstracts
Organizer: **(Tony) Jianguo Sun** (University of Missouri, USA) **Guanyu Hu** (Michigan State University, USA)	
Chair: **Guanyu Hu** (Michigan State University, USA)	
15:10-15:30	*Bayesian Causal Discovery with Cycles and Latent Confounders* **Yanxun Xu** (Johns Hopkins University, USA)
15:35-15:55	*Bayesian Automated Learning of Sparsity in Risk Prediction with Application to Whole-brain Functional Connectivity Analysis* **Xia Wang** (University of Cincinnati, USA)
16:00-16:20	Bayesian design and analysis methods for decentralized clinical trials **Ruitao Lin** (The University of Texas M.D. Anderson Cancer Center, USA)
16:25-16:45	*Bayesian Dose-Response Meta-Analysis for Predictive Biomarkers using Aggregate and Individual Participant Data with Data Augmentation for Precision Medicine* **J. Jack Lee** (The University of Texas M.D. Anderson Cancer Center, USA)

December 18 (Thursday):

Keynote Speech (III) [09:00 – 10:00]:

Location: S1
Title: ***Inference for Interpretable Machine Learning: Feature Importance and Beyond***
Speaker: **Genevera I. Allen** (Columbia University, USA)
Chair: **Hsin-Cheng Huang** (Academia Sinica, Taiwan, ROC)

Keynote Speech (IV) [10:20 – 11:20]:

Location: S1
Title: ***Shaping the Future: The Expanding Role of Pharma Statisticians in the AI Era***
Speaker: **Amy Xia** (Amgen Inc., USA)
Chair: **Xinping Cui** (University of California, Riverside, USA)

Parallel Sessions [Dec. 18, 12:50 – 14:30]:

18a1 - Recent Developments for Data Science	
Location: S1	To Abstracts
Organizer: **Henry Horng-Shing Lu** (Kaohsiung Medical University / National Yang Ming Chiao Tung University, Taiwan, ROC)	
Chair: **Rong Chen** (Rutgers University, USA)	
12:50-13:10	*AI, BI & SI—Artificial, Biological and Statistical Intelligences* **Dennis Lin** (Purdue University, USA)
13:15-13:35	*Solving the Mysteries of Place Cells and Grid Cells by Representation Learning* **Yingnian Wu** (University of California, Los Angeles, USA)
13:40-14:00	*Generate Diverse Protein Conformations through AlphaFold* **Samuel Kou** (Harvard University, USA)
14:05-14:25	*Recent Advances in MM Optimization Algorithms* **Hua Zhou** (University of California, Los Angeles, USA)

Back to Sessions List

Parallel Sessions [Dec. 18, 12:50 – 14:30]

18a2 - Statistical Innovations for High-Dimensional and Time-Dependent Data	
Location: S2	To Abstracts
Organizer: **Ching-Kang Ing** (National Tsing Hua University, Taiwan, ROC)	
Chair: **Ching-Kang Ing** (National Tsing Hua University, Taiwan, ROC)	
12:50-13:10	*Asymptotic FDR Control with Model-X Knockoffs: Is Moments Matching Sufficient?* **Yingying Fan** (University of Southern California, USA)
13:15-13:35	*Detection of Dynamic Instability by Dispersion Ratios in Local Block Lyapunov Exponent Diagrams* **Yan Liu** (Waseda University, Japan)
13:40-14:00	*Variable Selection for High-Dimensional Heteroscedastic Regression and Its Applications* **Hsueh-Han Huang** (Academia Sinica, Taiwan, ROC)
14:05-14:25	*Adaptive High-Dimensional Model Selection via Chebyshev's Greedy Algorithm* **Chi-Shian Dai** (National Cheng Kung University, Taiwan, ROC)

18a3 - Analysis of Complex Data	
Location: S3	To Abstracts
Organizer: **Ying Hung** (Rutgers University, USA)	
Chair: **Shihao Yang** (Georgia Institute of Technology, USA)	
12:50-13:10	*Statistical Thinking and AI Transformers: A Two-Way Exchange Between Time Series and Attention Mechanisms* **Shihao Yang** (Georgia Institute of Technology, USA)
13:15-13:35	*HeteroJIVE: Joint Subspace Estimation for Heterogeneous MultiView Data* **Jingyang Li** (University of Michigan, USA)
13:40-14:00	*A Riemannian Factor Model for Manifold-valued Time Series* **Shuo-Chieh Huang** (Rutgers University, USA)
14:05-14:25	*Enhancing Generalizability and Fairness of HIV Risk Predictions: A Machine Learning Approach Using EHR Data* **Hulin Wu** (University of Texas Health Science Center at Houston, USA)

Parallel Sessions [Dec. 18, 12:50 – 14:30]

18a5 - New Data Analytics for Evaluating Complex Associations	
Location: S5	To Abstracts
Organizer: **Peter Song** (University of Michigan, USA)	
Chair: **Hua Shen** (University of Calgary, Canada)	
12:50-13:10	*Examining Directional Association between Depression and Anxiety in US Medical Interns* **Soumik Purkayashta** (University of Pittsburgh, USA)
13:15-13:35	*Statistical Methods for Chemical Mixtures: A Roadmap for Practitioners Using Simulation Studies and a Sample Data Analysis in the PROTECT Cohort* **Wei Hao** (University of Michigan, USA)
13:40-14:00	*Supervised Fusion Learning of Physical Activity Features: Functional Frameworks and Longitudinal Analysis with L_0 Regularization* **Margaret Banker** (Northwestern University Feinberg School of Medicine, USA)
14:05-14:25	*Estimation and Inference of Quantile Spatially Varying Coefficient Models Over Complicated Domains* **Myungjin Kim** (Kyungpook National University, South Korea)

18a6 - Causal Inference for Survival Data	
Location: A1	To Abstracts
Organizer: **C. Jason Liang** (National Institute of Allergy and Infectious Diseases, USA)	
Chair: **C. Jason Liang** (National Institute of Allergy and Infectious Diseases, USA)	
12:50-13:10	*Leveraging External Individualized Prediction Models in Bayesian Survival Analysis* **Mi-Ok Kim** (University of California San Francisco, USA)
13:15-13:35	*Propensity Weighting Plus Adjustment in Proportional Hazards Model Is Not Doubly Robust* **Erin Gabriel** (University of Copenhagen, Denmark)
13:40-14:00	*Event History Regression with Pseudo-Observations: Computational Approaches and Causal Inference* **Michael Sachs** (University of Copenhagen, Denmark)
14:05-14:25	*Transporting Evidence from and to External Studies by Leveraging Aggregate Data* **Chiung-Yu Huang** (University of California, San Francisco, USA)

Parallel Sessions [Dec. 18, 12:50 – 14:30]

18a7 - Recent Advances in Time Series Analysis	
Location: A2	To Abstracts
Organizer: **George Michailidis** (University of California, Los Angeles, USA)	
Chair: **Junho Yang** (Academia Sinica, Taiwan, ROC)	
12:50-13:15	*Data Secure Transfer Learning from Heterogeneous Low Rank and Sparse Panel VAR Models* **George Michailidis** (University of California, Los Angeles, USA)
13:25-13:50	*Fast Segmentation of Watermarked Texts from Large Language Models through Epidemic Change-Points Framework* **Sayar Karmakar** (University of Florida, USA)
14:00-14:25	*Monitoring and Early Detection of Instability in Manufacturing Process Using Vector Autoregression Models* **Yi-Ting Wang** (National Taiwan University, Taiwan, ROC)

18a8 - 3D Protein Structure Informatics	
Location: A3	To Abstracts
Organizer: **I-Ping Tu** (Academia Sinica, Taiwan, ROC)	
Chair: **I-Ping Tu** (Academia Sinica, Taiwan, ROC)	
12:50-13:15	*Recent Advances of Structural Biology via Single Particle Cryogenic Electron Microscopy and the Remaining Challenges* **Wei-Hau Chang** (Academia Sinica, Taiwan, ROC)
13:25-13:50	*CRISP: A Modular Platform for Cryo-EM Image Segmentation and Processing with Conditional Random Field* **Szu-Chi Chung** (National Sun Yat-sen University, Taiwan, ROC)
14:00-14:25	*A Robust Hierarchical Linear Model for Cryo-EM Analysis* **I-Ping Tu** (Academia Sinica, Taiwan, ROC)

Parallel Sessions [Dec. 18, 12:50 – 14:30]

18a9 - New Developments in High-Dimensional Matrix and Network Analysis	
Location: A4	To Abstracts
Organizer: **Shujie Ma** (University of California, Riverside, USA)	
Chair: **Wanjie Wang** (National University of Singapore, Singapore)	
12:50-13:10	(TBD) **Wen Zhou** (New York University, USA)
13:15-13:35	*Factor Models of Matrix-Valued Time Series: Nonstationarity and Cointegration* **Degui Li** (University of Macau, Macau)
13:40-14:00	*Estimating SNR in High-Dimensional Linear Models* **Xiaodong Li** (University of California, Davis, USA)
14:05-14:25	*Natural Covariate-Adjusted Graphical Regression* **Guo Yu** (University of California, Santa Barbara, USA)

18a10 - Recent Methods Development in High-Dimensional Omics Studies	
Location: A5	To Abstracts
Organizer: **Hongyuan Cao** (Florida State University, USA)	
Chair: **Chien-Ming Chi** (Academia Sinica, Taiwan, ROC)	
12:50-13:10	*The Blurred Line between Genes and Environments: Insights from GWAS of Family Members' Phenotypes* **Qiongshi Lu** (University of Wisconsin-Madison, USA)
13:15-13:35	*Inferring Cell-Type-Specific Co-Methylation Networks from Single-Cell DNA Methylation Data* **Jiebiao Wang** (University of Pittsburgh, USA)
13:40-14:00	*Single-Cell Multiomic Analysis of Circadian Rhythmicity* **Yuchao Jiang** (Texas A&M University, USA)
14:05-14:25	*mist: A Hierarchical Bayesian Framework for Detecting Differential DNA Methylation Dynamics in Single-Cell Data* **Hao Feng** (The University of Texas Health Science Center at Houston, USA)

Parallel Sessions [Dec. 18, 14:50 – 16:30]:

18b1 - Spatial and Environmental Statistics	
Location: S1	To Abstracts
Organizer: **Hsin-Cheng Huang** (Academia Sinica, Taiwan, ROC)	
Chair: **Sylvia R. Esterby** (University of British Columbia, Canada)	
14:50-15:15	*CQUESST: A Bayesian Framework for Soil-Carbon Sequestration* **Noel Cressie** (University of Wollongong, Australia)
15:25-15:50	*Modelling Count Data in the Presence of Intervention* **Abdel El-Shaarawi** (Cairo University, Egypt)
16:00-16:25	*Covariate-Dependent Spatio-Temporal Covariance Models* **Nan-Jung Hsu** (National Tsing-Hua University, Taiwan, ROC)

18b2 - New Insights into Inference and Causality	
Location: S2	To Abstracts
Organizer: **Lei Liu** (Washington University in St. Louis, USA) **Ying Ding** (University of Pittsburgh, USA)	
Chair: **Wei Chen** (University of Pittsburgh, USA)	
14:50-15:10	*Model-Free Inference for Characterizing Protein Mutations through a Coevolutionary Lens* **Zhao Ren** (University of Pittsburgh, USA)
15:15-15:35	*Microbial Causal Mediation Analysis under Spatially Correlated Exposure* **Huilin Li** (New York University, USA)
15:40-16:00	*Testing Composite Null Hypotheses with High-Dimensional Dependent Data: A Computationally Scalable FDR-Controlling Procedure* **Hongyuan Cao** (Florida State University, USA)
16:05-16:25	*D-CDLF: Decomposition of Common and Distinctive Latent Factors for Multi-View High-Dimensional Data* **Hai Shu** (New York University, USA)

Parallel Sessions [Dec. 18, 14:50 – 16:30]

18b3 - Recent Developments for Biomedical Statistics	
Location: S3	To Abstracts
Organizer: **Henry Horng-Shing Lu** (Kaohsiung Medical University / National Yang Ming Chiao Tung University, Taiwan, ROC)	
Chair: **Xiao-Li Meng** (Harvard University, USA)	
14:50-15:10	*Longitudinal First Hitting-Time Models with Extension to Neural Network* **Mei-Ling Ting Lee** (University of Maryland, USA)
15:15-15:35	*Time-Dependent Pseudo R-Squared for Assessing Predictive Performance in Competing Risks Data* **Gang Li** (University of California, Los Angeles, USA)
15:40-16:00	*Assessing Transcriptomic Heterogeneity of Single-Cell RNASeq Data by Bulk-Level Gene Expression Data* **Chen-Hsiang Yeang** (Academia Sinica, Taiwan, ROC)
16:05-16:25	*A Unified Framework for Statistical Inference and Power Analysis of Single and Comparative F_β Scores* **Chih-Yuan Hsu** (Vanderbilt University Medical Center, USA)

18b4 - Recent Advances in Optimal Experimental Designs	
Location: S4	To Abstracts
Organizer: **Weng Kee Wong** (University of California, Los Angeles, USA) **Jesús López Fidalgo** (University of Navarra, Spain)	
Chair: **Jesús López Fidalgo** (University of Navarra, Spain)	
14:50-15:10	*Design Strategies for Robust Subsampling under Model Misspecification* **Carlos de la Calle** (Universidad de Oviedo, Spain)
15:15-15:35	*A Design Optimality Criterion Based on the AUC for Classification* **Jesús López Fidalgo** (University of Navarra, Spain)
15:40-16:00	*Models and Procedures for the Estimation of Blood Alcohol Concentration in the Human Body* **Juan Manuel Rodriguez-Diaz** (University of Salamanca, Spain)
16:05-16:25	*Swarm-Based Search Procedure for Finding Optimal Multi-Stage Designs for Phase II Clinical Trials* **Ping-Yang Chen** (National Taipei University, Taiwan, ROC)

Parallel Sessions [Dec. 18, 14:50 – 16:30]

18b5 - High-Dimensional Models with Applications in Biomedical Sciences	
Location: S5	To Abstracts
Organizer: **Yi Li** (University of Michigan, USA)	
Chair: **Hsueh-Han Huang** (Academia Sinica, Taiwan, ROC)	
14:50-15:10	*Inference on Deep Neural Network Estimators* **Yi Li** (University of Michigan, USA)
15:15-15:35	*Causal Learning with Label Noise: A Classification Approach for Paired Vectors* **Grace Yi** (University of Western Ontario, Canada)
15:40-16:00	*In-Sample Evaluation of Subgroups Identified by Generic Machine Learning* **XinZhou Guo** (The Hong Kong University of Science and Technology, Hong Kong)
16:05-16:25	*A Unified Framework of Analyzing Missing Data and Variable Selection Using Regularized Likelihood* **Wenqing He** (University of Western Ontario, Canada)

18b6 - Modern Machine Learning in the Big Data Era	
Location: A1	To Abstracts
Organizer: **Yichuan Zhao** (Georgia State University, USA)	
Chair: **Yichuan Zhao** (Georgia State University, USA)	
14:50-15:10	*Logistics Regression Model for Differentially-Private Matrix Masking Data* **Samuel S. Wu** (University of South Florida, USA)
15:15-15:35	*Partially-Global Fréchet Regression* **Yichao Wu** (University of Illinois Chicago, USA)
15:40-16:00	*Learning nonparametric graphical model on heterogeneous network-linked data* **Junhui Wang** (The Chinese University of Hong Kong, Hong Kong)
16:05-16:25	*Residual-Based Subdata Selection for Local Linear Regression and Its Extension to Partial Linear Models* **Chia-Wei Lin** (National Tsing Hua University, Taiwan, ROC)

Parallel Sessions [Dec. 18, 14:50 – 16:30]

18b7 - Classical Meets Cutting-Edge: Regression, Mixtures, and Joint Models in Biomedical Research	
Location: A2	To Abstracts
Organizer: **Xinping Cui** (University of California, Riverside, USA)	
Chair: **Xinping Cui** (University of California, Riverside, USA)	
14:50-15:10	*A Perturbation Subsampling for Large Scale Data* **Zhezhen Jin** (Columbia University, USA)
15:15-15:35	*Subgroup Mixture Challenges in Bridging Studies for Predictive Biomarkers* **Szu-Yu Tang** (Pfizer Inc., USA)
15:40-16:00	*Scalable Joint Modeling of Multiple Longitudinal Biomarkers and Competing Risks Time-to-Event Data: with Applications to MegaScale Health Research* **Xinping Cui** (University of California, Riverside, USA)
16:05-16:25	*Metaheuristics as a General-Purpose Optimization Tool for Statistical Research* **Weng Kee Wong** (University of California, Los Angeles, USA)

18b8 - Exploring Phenomena Through Mathematical Modeling	
Location: A3	To Abstracts
Organizer: **Masato Abe** (Doshisha University, Japan)	
Chair: **Masato Abe** (Doshisha University, Japan)	
14:50-15:10	*Mathematical Models of the Feedback between Population Dynamics and Biological or Cultural Evolution* **Shota Shibasaki** (Doshisha University, Japan)
15:15-15:35	*Structure-Plasticity Interactions Shape Self-Organized Criticality in Neural Networks* **Yoshiki A. Sugimoto** (Doshisha University, Japan)
15:40-16:00	*Explaining Parasite Diversity and Host Diversity in Nature* **Wei-Chung Liu** (Academia Sinica, Taiwan, ROC)
16:05-16:25	*Mapping Citation Tendencies among Broad Subject Groups: Closed Fields, Asymmetric Exchange, Unexpected Ties and Structural Disconnection* **Frederick Kin Hing Phoa** (Academia Sinica, Taiwan, ROC)

Parallel Sessions [Dec. 18, 14:50 – 16:30]

18b9 - Copula and Dependence Modeling	
Location: A4	To Abstracts
Organizer: **Takeshi Emura** (Hiroshima University, Japan)	
Chair: **Takeshi Emura** (Hiroshima University, Japan)	
14:50-15:15	*B-Spline Copula and Its Estimation* **Xiaoling Dou** (International Christian University, Japan)
15:25-15:50	*Measuring Multivariate Regression Association via Spatial Sign* **Jia-Han Shih** (National Sun Yat-sen University, Taiwan, ROC)
16:00-16:25	*The Trivariate Wrapped Cauchy Copula* **Shogo Kato** (Institute of Statistical Mathematics, Japan)

18b10 - Modern Bayesian Tools for Modeling and Inference	
Location: A5	To Abstracts
Organizer: **Haiyan Huang** (University of California, Berkeley, USA) **Feng Liang** (University of Illinois at Urbana-Champaign, USA)	
Chair: **Yan-Bin Chen** (National Taiwan University, Taiwan, ROC)	
14:50-15:10	*A Bayesian Estimator of Sample Size* **Yuan Ji** (University of Chicago, USA)
15:15-15:35	*Sampling from the Random Linear Model via Stochastic Localization Up to the AMP Threshold* **Jingbo Liu** (University of Illinois, USA)
15:40-16:00	*Bayesian Smoothing and Feature Selection via Variational Automatic Relevance Determination* **Feng Liang** (University of Illinois at Urbana-Champaign, USA)
16:05-16:25	*Prediction Interval Transfer Learning for Linear Regression using an Empirical Bayes Approach* **Min Zhang** (University of California, Irvine, USA)

December 19 (Friday):

ICSA Pao-Lu Hsu Award (I) [09:00 – 10:00]:

Location: S1

Title: ***High-Dimensional Clustering via a Latent Transformation Mixture Model***

Speaker: **Hui Zou** (University of Minnesota, USA)

Chair: **Ying Zhang** (University of Nebraska Medical Center, USA)

Parallel Sessions [Dec. 19, 10:20 – 12:00]:

19a1 - Statistica Sinica Special Invited Papers	
Location: S1	To Abstracts
Organizer: **Huixia Judy Wang** (Rice University, USA)	
Chair: **Huixia Judy Wang** (Rice University, USA)	
1	*The Method of Limits and Its Application to The Analysis of Count Data in Genome-Wide Association Studies* **Jiming Jiang** (University of California, Davis, USA)
2	*Weighted Conditional Network Testing for Multiple High-Dimensional Correlated Data Sets* **Inyoung Kim** (Virginia Tech, USA)

19a2 - Frontiers in Statistical Inference: Dependence and Data Privacy	
Location: S2	To Abstracts
Organizer: **Yingying Fan** (University of Southern California, USA)	
Chair: **Yingying Fan** (University of Southern California, USA)	
10:20-10:40	*Statistical Inference for Differentially Private Stochastic Gradient* **Zhanrui Cai** (The University of Hong Kong, Hong Kong)
10:45-11:05	*Greedy Model Selection under Sparsity and Covariate Shift* **Ching-Kang Ing** (National Tsing Hua University, Taiwan, ROC)
11:10-11:30	*Tying Maximum Likelihood Estimation for Dependent Data* **Qingfeng Liu** (Hosei University, Japan)
11:35-11:55	*LLM-Powered Prediction Inference with Online Text Time Series* **Jinchi Lv** (University of Southern California, USA)

Parallel Sessions [Dec. 19, 10:20 – 12:00]

19a3 - Advances in Causality, Reinforcement Learning, and Business Analytics	
Location: S3	To Abstracts
Organizer: **Xiaowu Dai** (University of California, Los Angeles, USA)	
Chair: **Chien-Ming Chi** (Academia Sinica, Taiwan, ROC)	
10:20-10:40	*Learning Robust Decision Rules for Censored and Confounded Data* **Yifan Cui** (Zhejiang University, China)
10:45-11:05	*Causal Inference for All: Marginal Causal Effects for Outcomes Truncated by Death* **Linbo Wang** (University of Toronto, Canada)
11:10-11:30	*Quantum speedups for multiproposal MCMC* **Andrew J. Holbrook** (University of California, Los Angeles, USA)
11:35-11:55	*Indirect Statistical Inference with Guaranteed Necessity and Sufficiency* **Zhengjun Zhang** (University of the Chinese Academy of Sciences, China)

Back to Sessions List

Parallel Sessions [Dec. 19, 10:20 – 12:00]

19a4 - Advanced Methods for Novel Biomedical Data Types	
Location: S4	To Abstracts
Organizer: **Cheng Zheng** (University of Nebraska Medical Center, USA)	
Chair: **Ming-Yueh Huang** (Academia Sinica, Taiwan, ROC)	
10:20-10:40	*BrainGeneBot: A GPT-engineered, User-Driven Genetic Data Exploration with Polygenic Risk Scores Ranking in Alzheimer's Disease* **Zhongming Zhao** (University of Texas Health Science Center at Houston, USA)
10:45-11:05	*High-dimensional Markov-switching Ordinary Differential Processes* **Katherine Tsai** (Apple, USA)
11:10-11:30	*Controlling False Discover Rate for High Dimensional Mediator Selection in Non-linear Models* **Ran Dai** (University of Nebraska Medical Center, USA)
11:35-11:55	*When Few Labeled Target Data Suffice: A Theory of Semi-Supervised Domain Adaptation via Fine-Tuning from Multiple Adaptive Starts* **Wooseok Ha** (Korea Advanced Institute of Science & Technology, South Korea)

19a5 - Recent Advancements in Network and Correlated Data Analysis	
Location: S5	To Abstracts
Organizer: **Shujie Ma** (University of California, Riverside, USA)	
Chair: **Hsueh-Han Huang** (Academia Sinica, Taiwan, ROC)	
10:20-10:40	(TBD) **Weining Wang** (University of Bristol, UK)
10:45-11:05	*Covariance-Based Clustering and Biclustering via the Heterogeneous Block Covariance Model and Variants* **Yunpeng Zhao** (Colorado State University, USA)
11:10-11:30	*Adaptive Block-Based Change-Point Detection for Sparse Spatially Clustered Data with Applications in Remote Sensing Imaging* **Lynna Chu** (Iowa State University, USA)
11:35-11:55	*Graph Release with Assured Node Differential Privacy* **Tianxi Li** (University of Minnesota, USA)

Parallel Sessions [Dec. 19, 10:20 – 12:00]

19a6 - Deep Learning and Artificial Intelligence	
Location: A1	To Abstracts
Organizer: **Lei Liu** (Washington University in St. Louis, USA) **Ying Ding** (University of Pittsburgh, USA)	
Chair: **Menggang Yu** (University of Michigan, USA)	
10:20-10:40	*Advancing Responsible Statistical and AI/ML Methods for Harnessing the Power of Electronic Health Records* **Qi Long** (University of Pennsylvania, USA)
10:45-11:05	*Deep Survival Analysis for Competing Risk Modeling with Functional Covariates and Missing Data Imputation* **Xiaofeng Wang** (Cleveland Clinic, USA)
11:10-11:30	*Mini-Batch Estimation for Deep Cox Models: Statistical Foundations and Practical Guidance* **Ying Ding** (University of Pittsburgh, USA)
11:35-11:55	*A Deep Learning Feature Importance Test for Integrating Informative High-dimensional Biomarkers* **Baiming Zou** (University of North Carolina at Chapel Hill, USA)

19a7 - Recent Developments on Biostatistics and AI	
Location: A2	To Abstracts
Organizer: **Xiaotong Shen** (University of Minnesota, USA)	
Chair: **An-Shun Tai** (National Tsing Hua University, Taiwan, ROC)	
10:20-10:45	*Latent Noise Injection for Private and Statistically Aligned Synthetic Data Generation* **Lu Tian** (Stanford University, USA)
10:55-11:20	*MR2G: A Novel Framework for Causal Network Inference Using GWAS Summary Data* **Haoran Xue** (City University of Hong Kong, Hong Kong)
11:30-11:55	*Modeling Non-Uniform Hypergraphs Using Determinantal Point Processes* **Ji Zhu** (University of Michigan, USA)

Parallel Sessions [Dec. 19, 10:20 – 12:00]

19a8 - Recent Advances in Clinical Trials	
Location: A3	To Abstracts
Organizer: **Guei-Feng (Cindy) Tsai** (Center for Drug Evaluation, Taiwan, ROC)	
Chair: **Chin-Fu Hsiao** (National Health Research Institutes, Taiwan, ROC)	
10:20-10:40	*Patient-Centric Pragmatic Trials: Opening the DOOR to Benefit: Risk-Based Evaluation* **Toshimitsu Hamasaki** (The George Washington University, USA)
10:45-11:05	*Adaptive Population Selection Designs for Clinical Trials with Multiple Endpoints* **Koko Asakura** (National Cerebral and Cardiovascular Center, Japan)
11:10-11:30	*Sample Size Assessment for Survival Trial Designs with Covariate-Adaptive Randomization* **Pei-Fang Su** (National Cheng Kung University, Taiwan, ROC)
11:35-11:55	*Comparing MCP-MOD and Ordinal Linear Contrast Test in Dose Finding Clinical Trials: A Thorough Examination* **Naitee Ting** (StatsVita, LLC, USA)

19a9 - Recent Advances in Statistics and AI	
Location: A4	To Abstracts
Organizer: **Jane-Ling Wang** (University of California, Davis, USA)	
Chair: **Ciren Jiang** (National Taiwan University, Taiwan, ROC)	
10:20-10:40	*Simultaneous Clustering and Estimation of Additive Shape Invariant Models for Recurrent Event Data* **Shizhe Chen** (University of California, Davis, USA)
10:45-11:05	*Multiview Manifold Learning for High-Dimensional and Noisy Data Analysis* **Xiucai Ding** (University of California, Davis, USA)
11:10-11:30	*Quantile Small Area Estimation via Singh-Maddala Mixed Model Prediction* **Thuan Nguyen** (Oregon Health and Science University, USA)
11:35-11:55	*Interpretable Transformer Regression for Functional and Longitudinal Covariates* **Cynthia Juang** (University of California, Davis, USA)

Parallel Sessions [Dec. 19, 10:20 – 12:00]

19a10 - Modern Bayesian and Machine Learning Methods for Precision Medicine and Digital Health	
Location: A5	To Abstracts
Organizer: **Yuanjia Wang** (Columbia University, USA)	
Chair: **Yu Gu** (The University of Hong Kong, Hong Kong)	
10:20-10:40	*Joint Mixed Membership Modeling of Multivariate Longitudinal and Survival Data* **Xinyuan Song** (The Chinese University of Hong Kong, Hong Kong)
10:45-11:05	*Building a Dose Toxo-equivalence Model from a Bayesian Meta-analysis of Published Clinical Trials* **Qingxia Chen** (Vanderbilt University Medical Center, USA)
11:10-11:30	*Quasi Instrumental Variable Methods for Stable Hidden Confounding and Binary Outcome* **Zhonghua Liu** (Columbia University, USA)
11:35-11:55	*Clustering-Informed Shared-Structure Variational Autoencoder for Missing Data Imputation in Large-Scale Healthcare Data* **Yuan Chen** (Memorial Sloan Kettering Cancer Center, USA)

Back to Sessions List

Parallel Sessions [Dec. 19, 12:50 – 14:30]:

19b1 - What is Obscure about Random Objects?	
Location: S1	To Abstracts
Organizer: **Jane-Ling Wang** (University of California, Davis, USA)	
Chair: **Jeng-Min Chiou** (Academia Sinica, Taiwan, ROC)	
12:50-13:15	*Modeling Amplitude and Phase Variation of Multivariate Random Processes in Geodesic Spaces* **Yaqing Chen** (Rutgers University, USA)
13:25-13:50	*Transfer Learning for Functional Linear Regression* **Zhenhua Lin** (National University of Singapore, Singapore)
14:00-14:25	*Inference for Dispersion and Curvature of Random Objects* **Hans-Georg Mueller** (University of California, Davis, USA)

19b2 - Data Visualization	
Location: S2	To Abstracts
Organizer: **Juergen Symanzik** (Utah State University, USA)	
Chair: **Chun-Houh Chen** (Academia Sinica, Taiwan, ROC)	
12:50-13:10	*Improving Interpretability in Machine Learning Using Confidence Intervals in ALE Plots* **John Stevens** (Utah State University, USA)
13:15-13:35	*Guided Data Visualization via Random Forests and Manifold Learning* **Kevin Moon** (Utah State University, USA)
13:40-14:00	*AI-assisted Data Visualization and Analytic* **Kwan-Liu Ma** (University of California, Davis, USA)
14:05-14:25	*iISOMAP: Nonlinear Dimensionality Reduction and Visualization for Interval-Valued Data via Geodesic Distance Preservation* **Han-Ming (Hank) Wu** (National Chengchi University, Taiwan, ROC)

Parallel Sessions [Dec. 19, 12:50 – 14:30]

19b3 - Recent Advances in Nonparametric Methods and Their Applications	
Location: S3	To Abstracts
Organizer: **Yichuan Zhao** (Georgia State University, USA)	
Chair: **Yichuan Zhao** (Georgia State University, USA)	
12:50-13:10	*Bivariate Analysis of Distribution Functions Under Biased Sampling* **Hsin-Wen Chang** (Academia Sinica, Taiwan, ROC)
13:15-13:35	*Regression in 2-Wasserstein Distance* **Li-Shan Huang** (National Tsing Hua University, Taiwan, ROC)
13:40-14:00	*Debiased Inference for High-Dimensional Censored Quantile Regression* **Tony Sit** (The Chinese University of Hong Kong, Hong Kong)
14:05-14:25	*Data Integration in Survey Sampling and Official Statistics* **Changbao Wu** (University of Waterloo, Canada)

19b4 - Extending Canonical Conformal Predictions to Meet the Practitioners' Needs	
Location: S4	To Abstracts
Organizer: **Nicolas Brunel** (ENSIIE & University Paris-Saclay, France)	
Chair: **Tso-Jung Yen** (Academia Sinica, Taiwan, ROC)	
12:50-13:10	*Mask-Conditional Conformal Prediction: Valid Uncertainty for All Missing Data Mechanisms* **Jiarong Fan** (University Paris-Saclay, France)
13:15-13:35	*Robust Conformal Prediction Using Privileged Information* **Shai Feldman** (Technion, Israel)
13:40-14:00	*Adaptive Coverage Policies in Conformal Prediction* **Etienne Gauthier** (University Inria, France)
14:05-14:25	*Towards a Rigorous Evaluation of RAG Systems: The Challenge of Due Diligence* **Nicolas Brunel** (ENSIIE & University Paris Saclay, France)

Parallel Sessions [Dec. 19, 12:50 – 14:30]

19b5 - Recent Research in Statistical Process Control, Part II	
Location: S5	To Abstracts
Organizer: **Xiulin Xie** (Florida State University, USA) **Peihua Qiu** (University of Florida, USA)	
Chair: **Peihua Qiu** (University of Florida, USA)	
12:50-13:15	*Univariate Self-Starting Shiryaev (3S): A Bayesian Change Point Model for Online Monitoring of Short Runs* **Panagiotis Tsiamyrtzis** (Politecnico di Milano, Italy)
13:25-13:50	*Adaptive Sampling in Profile Monitoring Through Bandits* **Chen Nan** (National University of Singapore, Singapore)
14:00-14:25	*Multivariate Control Charts for Correlated Quality Variables of Different Types* **Bai-Yau Yeh** (Bowling Green State University, USA)

19b6 - Recent Developments in AI	
Location: A1	To Abstracts
Organizer: **Hongyuan Cao** (Florida State University, USA)	
Chair: **Hongyuan Cao** (Florida State University, USA)	
12:50-13:15	*Privacy-Preserving LLM Alignment via Private Reward Modeling: A Holistic and Data-Efficient Framework* **Young Hyun Cho** (Purdue University, USA)
13:25-13:50	*Bridging Spatial Transcriptomics and Histopathology through AI* **Wei Chen** (University of Pittsburgh, USA)
14:00-14:25	*Fair Graph Learning Without Complete Demographics* **Fang Liu** (University of Notre Dame, USA)

Parallel Sessions [Dec. 19, 12:50 – 14:30]

19b7 - Recent Developments of Statistical Methods in Case-Cohort Study Design and Dependent Sampling	
Location: A2	To Abstracts
Organizer: **Chun-Shu Chen** (National Central University, Taiwan, ROC) **Feng-Chang Lin** (University of North Carolina at Chapel Hill, USA)	
Chair: **Feng-Chang Lin** (University of North Carolina at Chapel Hill, USA)	
12:50-13:10	*Super Learner for Survival Prediction in Case-Cohort and Generalized Case-Cohort Studies* **Jianwen Cai** (University of North Carolina at Chapel Hill, USA)
13:15-13:35	*Efficient Case-Cohort Design Using Balanced Sampling* **Sangwook Kang** (Yonsei University, South Korea)
13:40-14:00	*Improving Efficiency of Risk Prediction with Subsampled Cohort Data* **Yei Eun Shin** (Seoul National University, South Korea)
14:05-14:25	*Semiparametric Regression Analysis of Case-Cohort Studies with Multiple Interval-Censored Disease Outcomes* **Haibo Zhou** (University of North Carolina at Chapel Hill, USA)

19b8 - Recent Research Developments in Neuroimaging Data Analysis	
Location: A3	To Abstracts
Organizer: **Tingting Zhang** (University of Pittsburgh, USA)	
Chair: **Chen-Hsiang Yeang** (Academia Sinica, Taiwan, ROC)	
12:50-13:15	*Choice of Metric and the Effect of Scan Length for Reliability in Resting-State fMRI* **Todd R. Ogden** (Columbia University, USA)
13:25-13:50	*Clarifying and Extending Permutation Tests on Brain Map Correspondence Through Mixed-Effects Modeling* **Tingting Zhang** (University of Pittsburgh, USA)
14:00-14:25	*Threshold Spatial Attention Transformer for Efficient Image Generation* **Jian Kang** (University of Michigan, USA)

Parallel Sessions [Dec. 19, 12:50 – 14:30]

19b9 - Causal Inference: Episode I	
Location: A4	To Abstracts
Organizer: **Peng Ding** (University of California, Berkeley, USA)	
Chair: **Ming-Yueh Huang** (Academia Sinica, Taiwan, ROC)	
12:50-13:15	*Matching for Causal Inference* **Fang Han** (University of Washington, USA)
13:25-13:50	*Synthetic Nearest Neighbours: Extending Synthetic Controls for Matrix Completion with Missing Not at Random Data* **Dennis Shen** (University of Southern California, USA)
14:00-14:25	*A Pleiotropy-free Bayesian Model for Two- Sample Summary-Data Mendelian Randomization with Binary Outcomes* **An-Shun Tai** (National Tsing Hua University, Taiwan, ROC)

19b10 - Recent Developments in Survival Analysis and Clinical Trial Methodology	
Location: A5	To Abstracts
Organizer: **Ying Lu** (Stanford University, USA)	
Chair: **Lu Tian** (Stanford University, USA)	
12:50-13:10	*Survival Data Analysis using Average Hazard with Survival Weight* **Hajime Uno** (Harvard Medical School / Dana-Farber Cancer Institute, USA)
13:15-13:35	*Tools for Randomized Clinical Trials Using Restricted Mean Survival Time and Average Hazard* **Miki Horiguchi** (Dana-Farber Cancer Institute, USA)
13:40-14:00	*Time-to-Event Analysis with Treatment Switches in Clinical Trials* **Jie Chen** (Taimei Intelligence Biopharma R&D, China)
14:05-14:25	*Missing Data Strategies for Generalized Pairwise Comparisons in Randomized Clinical Trials* **Ying Lu** (Stanford University, USA)

December 20 (Saturday):

Parallel Sessions [Dec. 20, 08:40 – 10:20]:

20a1 – Machine Learning / AI	
Location: S1	To Abstracts
Organizer: **Hua Tang** (Stanford University, USA) **Nancy R. Zhang** (University of Pennsylvania, USA)	
Chair: **Hua Tang** (Stanford University, USA)	
08:40-09:00	*Modeling and Predicting Single-Cell Multi-Gene Perturbation Responses* **Hongyu Zhao** (Yale University, USA)
09:05-09:25	*Transcriptomic Analysis and Image-Based Deep Learning Prognostic Model for Lung Adenocarcinoma* **Hsuan-Yu Chen** (Academia Sinica, Taiwan, ROC)
09:30-09:50	*Modeling the Impact of Personal Genome Variation on Molecular Phenotypes* **Nilah Monnier Ioannidis** (University of California, Berkeley, USA)
09:55-10:15	*An AI System to Help Scientists Write Expert-Level Empirical Software* **Marc Coram** (Google Zurich, Switzerland)

Back to Sessions List

Parallel Sessions [Dec. 20, 08:40 – 10:20]

20a2 - Recent Advancements in Design of Experiments	
Location: S2	To Abstracts
Organizer: **John Stufken** (George Mason University, USA)	
Chair: **John Stufken** (George Mason University, USA)	
08:40-09:00	*Optimal Designs for Network Experimentation with Unstructured Treatments* **Jing-Wen Huang** (Academia Sinica, Taiwan, ROC)
09:05-09:25	*Optimal Experimental Designs for Low-Rank Function Completion* **Ming-Hung (Jason) Kao** (Arizona State University, USA)
09:30-09:50	*Efficient Bayesian Estimation and Inference for Shapley Value via Optimal Design* **Wei Zheng** (University of Tennessee, USA)
09:55-10:15	*Cyclic SOAs and Moving Window Criteria for Space-Filling Designs* **Cheng-Yu Sun** (National Tsing Hua University, Taiwan, ROC)

20a3 - Causal Inference	
Location: S3	To Abstracts
Organizer: **Yen-Tsung Huang** (Academia Sinica, Taiwan, ROC)	
Chair: **Yen-Tsung Huang** (Academia Sinica, Taiwan, ROC)	
08:40-09:00	*Two-stage Adaptive Testing of Large-Scale Mediation Hypotheses* **Gary Chan** (University of Washington, USA)
09:05-09:25	*Leveraging Multi-Study, Multi-Outcome Data to Improve External Validity and Efficiency of Clinical Trials for Medications for Opioid Use Disorder* **Caleb Miles** (Columbia University, USA)
09:30-09:50	*Mediation Analysis with Graph Mediator* **Yi Zhao** (Indiana University School of Medicine, USA)
09:55-10:15	*Toward Flexible and Efficient Counterfactual Density Estimation* **Kwangho Kim** (Korea University, Korea)

Parallel Sessions [Dec. 20, 08:40 – 10:20]

20a4 - Bayesian Adaptive Designs for Oncology Dose Optimization	
Location: S4	To Abstracts
Organizer: **Fangrong Yan** (China Pharmaceutical University, China)	
Chair: **Junho Yang** (Academia Sinica, Taiwan, ROC)	
1	(TBD) **Mengyi Lu** (Nanjing Medical University, China)
2	(TBD) **Xin Chen** (China Pharmaceutical University, China)

20a5 - Recent Advances in Network and Tensor Data Analysis	
Location: S5	To Abstracts
Organizer: **Emma Jingfei Zhang** (Emory University, USA)	
Chair: **Biao Cai** (City University of Hong Kong, Hong Kong)	
08:40-09:00	*Learning and Inference for Low-Rank Models* **Dong Xia** (The Hong Kong University of Science and Technology, Hong Kong)
09:05-09:25	*Federated Community Detection in Bipartite Networks from Various Platforms* **Wanjie Wang** (National University of Singapore, Singapore)
09:30-09:50	*Online Tensor Inference* **Yichen Zhang** (Purdue University, USA)
09:55-10:15	*Generalized Tensor Completion with Non-Random Missingness* **Emma Jingfei Zhang** (Emory University, USA)

Parallel Sessions [Dec. 20, 08:40 – 10:20]

20a6 - Advances in Statistical Modelling and Inference for High-Dimensional and Functional Data	
Location: A1	To Abstracts
Organizer: **Jin-Ting Zhang** (National University of Singapore, Singapore)	
Chair: **Li-Shan Huang** (National Tsing Hua University, Taiwan, ROC)	
08:40-09:05	*Diffusion Models for High Dimensional Digital Twins* **Xin Tong** (National University of Singapore, Singapore)
09:15-09:40	*Two-Sample Tests for Equal Distributions in Separable Metric Spaces: A Unified Semimetric-Based Approach* **Jin-Ting Zhang** (National University of Singapore, Singapore)
09:50-10:15	*A Fast and Accurate Kernel-Based Independence Test with Applications to High-Dimensional and Functional Data* **Tianming Zhu** (National Institute of Education, Nanyang Technological University, Singapore)

20a7 - Recent Advances in Clinical Trials	
Location: A2	To Abstracts
Organizer: **Naitee Ting** (StatsVita, LLC, USA)	
Chair: **Naitee Ting** (StatsVita, LLC, USA)	
08:40-08:55	*Bayesian Extrapolation Design: Exposure-Response Curve Comparison between Pediatric and Adult Populations* **Jingjing Ye** (BeOne Medicines)
09:00-09:15	*Bayesian Inference for Cluster-Randomized Trials with Multivariate Outcomes Subject to Both Truncation by Death and Missingness* **Guangyu Tong** (Yale University, USA)
09:20-09:35	*Design of a Dual Randomized Trial in a Type 2 Hybrid Effectiveness-Implementation Study* **Feng-Chang Lin** (University of North Carolina at Chapel Hill, USA)
09:40-09:55	*ProjectDRIVE: Effect of In-Vehicle Feedback with and without Parent Communication Training on Teen Driving Behaviours* **Jingzhen (Ginger) Yang** (Nationwide Children's Hospital, USA)
10:00-10:15	*Disease Progression Modeling with Informative Censoring* **Tom Jensen** (Utah State University, USA)

Parallel Sessions [Dec. 20, 08:40 – 10:20]

20a8 - New Advances in Biostatistics and Bioinformatics	
Location: A3	To Abstracts
Organizer: **Shuangge Ma** (Yale University, USA) **Hao Mei** (Renmin University of China, China)	
Chair: **Shuangge Ma** (Yale University, USA) **Hao Mei** (Renmin University of China, China)	
08:40-09:00	*Change Surface Regression for Nonlinear Subgroup Identification* **Jialiang Li** (National University of Singapore, Singapore)
09:05-09:25	*Unraveling the Neurogenetic Architecture of Nicotine Use and Depression: A Network-Driven Integration of Brain Transcriptomes and GWAS* **Bao-Zhu Yang** (Yale University, USA)
09:30-09:50	*Enhancing Model Generalizability in Medical AI: Systematic Comparison of Categorical Encoding and Sampling Techniques for Imbalanced Data* **Chien-Wei Chuang** (Fu Jen Catholic University, Taiwan, ROC)
09:55-10:15	*Time-Varying Latent Space Modeling for Outcome-Based Human Disease Network* **Hao Mei** (Renmin University of China, China)

20a9 - Innovations in Statistical Methodology for Complex Data: Spatial Omics, Stochastic Optimization, Network Integration, and Microscopy Image Analysis	
Location: A4	To Abstracts
Organizer: **Yuchao Jiang** (Texas A&M University, USA)	
Chair: **Yuchao Jiang** (Texas A&M University, USA)	
08:40-09:00	*AI-powered Bayesian Methods for Analyzing Spatial Omics Data* **Qiwei Li** (The University of Texas at Dallas, USA)
09:05-09:25	*Covariate-Assisted Graph Matching* **Jesús Arroyo Relión** (Texas A&M University, USA)
09:30-09:50	(TBD) **Abhishek Roy** (Texas A&M University, USA)
09:55-10:15	*Seeing is Believing: Challenges and Opportunities for Super-Resolution Microscopy Image Data Analysis for Quantitative Molecular Biology* **Chongzhi Zang** (University of Virginia, USA)

Parallel Sessions [Dec. 20, 08:40 – 10:20]

20a10 - Statistical and Algorithmic Foundations of Diffusion Models	
Location: A5	To Abstracts
Organizer: **Yuxin Chen** (University of Pennsylvania, USA)	
Chair: **Yan-Bin Chen** (National Taiwan University, Taiwan, ROC)	
08:40-09:05	*Optimal Score Estimation via Empirical Bayes Smoothing* **Yihong Wu** (Yale University, USA)
09:15-09:40	*Provable Statistical and Computational Efficiency of Diffusion Models* **Changxiao Cai** (University of Michigan, USA)
09:50-10:15	*Transformers Provably Learn Chain-of-Thought Reasoning with Length Generalization* **Yuejie Chi** (Yale University, USA)

Back to Sessions List

Parallel Sessions [Dec. 20, 10:40 – 12:20]:

20b1 - Modern Functional Data Methods with Applications to Environmental Studies	
Location: S1	To Abstracts
Organizer: **Ting-Li Chen** (Academia Sinica, Taiwan, ROC)	
Chair: **Shih-Hao Huang** (National Central University, Taiwan, ROC)	
10:40-11:05	*Mean Shift for Clustering Functional Data: A Scalable Algorithm and Convergence Analysis* **Toshinari Morimoto** (National Taiwan University, Taiwan, ROC)
11:15-11:40	*Personalized Functional Principal Component Analysis with Applications* **Ruey S. Tsay** (National Tsing Hua University, Taiwan, ROC)
11:50-12:15	*Nonstationary Gaussian Scale Mixture Models for Spatial Functional Extremes* **Yen-Shiu Chin** (Academia Sinica, Taiwan, ROC)

20b2 - Subsampling	
Location: S2	To Abstracts
Organizer: **John Stufken** (George Mason University, USA)	
Chair: **Ming-Hung (Jason) Kao** (Arizona State University, USA)	
10:40-11:00	*Optimal Subdata Selection for Large-Scale Linear Regression Under Model Misspecification* **Min Yang** (University of Illinois at Chicago, USA)
11:05-11:25	*Influence-Guided Active Subsampling for High-Dimensional Ridge Regression with Application in GWAS* **Lin Wang** (Purdue University, USA)
11:30-11:50	*Robust Data Fusion via Subsampling* **HaiYing Wang** (University of Connecticut, USA)
11:55-12:15	*Efficient Subdata Selection for Parameter Estimation* **John Stufken** (George Mason University, USA)

Parallel Sessions [Dec. 20, 10:40 – 12:20]

20b3 - Recent Advances in Time Series Analysis	
Location: S3	To Abstracts
Organizer: **George Michailidis** (University of California, Los Angeles, USA)	
Chair: **George Michailidis** (University of California, Los Angeles, USA)	
10:40-11:05	*Advances in Spatial Integer-Valued Time Series Modeling* **Cathy Chen** (Feng Chia University, Taiwan, ROC)
11:15-11:40	*Granger Causality Tests for High Dimensional VAR Processes* **Ying Chao Hung** (National Taiwan University, Taiwan, ROC)
11:50-12:15	*Autotune: Fast, Efficient, and Automatic Tuning Parameter Selection for LASSO* **Sumanta Basu** (Cornell University, USA)

20b5 - Recent Developments in Survival Analysis and Deep Learning	
Location: S5	To Abstracts
Organizer: **Xingqiu Zhao** (The Hong Kong Polytechnic University, Hong Kong)	
Chair: **Kin Yat Liu** (The Chinese University of Hong Kong, Hong Kong)	
10:40-11:00	*Improving Inference and Variable Selection for Two-Phase Studies with High-Dimensional Covariates* Kin Yau Wong (The Hong Kong Polytechnic University, Hong Kong)
11:05-11:25	*Efficient Estimation for Functional Accelerated Failure Time Model* **Kin Yat Liu** (The Chinese University of Hong Kong, Hong Kong)
11:30-11:50	*Semiparametric Causal Inference for Right-Censored Outcomes with Many Weak Invalid Instruments* **Qiushi Bu** (City University of Hong Kong, Hong Kong)
11:55-12:15	*Efficient Estimation for Deep Accelerated Failure Time Model with Application to Credit Risk Analysis* **Kun Ren** (City University of Hong Kong, Hong Kong)

Parallel Sessions [Dec. 20, 10:40 – 12:20]

20b6 - Recent Advancements in Deep Learning and Graphical Models, and Model Selection	
Location: A1	To Abstracts
Organizer: **Jian Huang** (The Hong Kong Polytechnic University, Hong Kong)	
Chair: **Jing-Wen Huang** (Academia Sinica, Taiwan, ROC)	
10:40-11:05	*Nonparametric GARCH: A Deep Learning Approach* **Guohao Shen** (The Hong Kong Polytechnic University, Hong Kong)
11:15-11:40	*Learning Summary Statistic for Likelihood-Free Inference* **Rong Tang** (The University of Science and Technology, Hong Kong)
11:50-12:15	*A Framework for Comprehensive Model and Variable Selection* **Xiyue Liao** (San Diego State University, USA)

20b7 - Applied Probability	
Location: A2	To Abstracts
Organizer: **I-Ping Tu** (Academia Sinica, Taiwan, ROC)	
Chair: **Yi-Ching Yao** (Academia Sinica, Taiwan, ROC)	
10:40-11:05	*Limiting Spectral Distribution of Stochastic Block Model* **Mei-Hui Guo** (National Sun Yat-sen University, Taiwan, ROC)
11:15-11:40	*On the Limiting Properties of Empirical Spectral Distributions in Community-structured Networks* **May-Ru Chen** (National Sun Yat-sen University, Taiwan, ROC)
11:50-12:15	(TBD) **Shoou-Ren Hsiau** (National Changhua University of Education, Taiwan, ROC)

Parallel Sessions [Dec. 20, 10:40 – 12:20]

20b8 - Regulatory Advances in Clinical Trials	
Location: A3	To Abstracts
Organizer: **Guei-Feng (Cindy) Tsai** (Center for Drug Evaluation, Taiwan, ROC)	
Chair: **Guei-Feng (Cindy) Tsai** (Center for Drug Evaluation, Taiwan, ROC)	
10:40-11:00	*Recent Initiatives of PMDA: Highlights and Selected Review Cases* **Takumi Aoki** (Pharmaceuticals and Medical Devices Agency (PMDA), Japan)
11:05-11:25	*Statistical Methods and Regulatory Considerations in Phase I Clinical Trials* **Tzu-Chuan Lin** (Division of New Drugs, Center for Drug Evaluation, Taiwan, ROC)
11:30-11:50	*Adaptive Designs for Clinical Trials: Principles and Recommendations from the Draft ICH E20 Guideline* **Frank Bretz** (Novartis Pharma AG, Switzerland)
11:55-12:15	*Evolving Regulatory Statistical Considerations in Drug Development and Evaluation* **Qian Helen Li** (StatsVita, LLC, USA)

Back to Sessions List

Parallel Sessions [Dec. 20, 13:20 – 15:00]:

20c1 - New Fronts on Machine Learning	
Location: S1	To Abstracts
Organizer: **Jane-Ling Wang** (University of California, Davis, USA)	
Chair: **Jane-Ling Wang** (University of California, Davis, USA)	
13:20-13:45	*Deep Kernel Aalen-Johansen Estimator: An Interpretable and Flexible Neural Net Framework for Competing Risks* **George Chen** (Carnegie Mellon University, USA)
13:55-14:20	*Quantifying Uncertainty in Answers from any Language Model and Enhancing their Trustworthiness* **Jonas Mueller** (Cleanlab, USA)
14:30-14:55	*Accessing the Impact of Data Alteration: When Can R-Squared from Synthetic (or Corrupted) Data Be Trusted?* **Xiao-Li Meng** (Harvard University, USA)

20c2 - Modern Advances in Learning, Robust Modeling, and Structure for High-Dimensional Data	
Location: S2	To Abstracts
Organizer: **Su-Yun Huang** (Academia Sinica, Taiwan, ROC)	
Chair: **Shao-Hsuan Wang** (National Central University, Taiwan, ROC)	
13:20-13:40	*On the Asymptotic Properties of Product-PCA under the High-Dimensional Setting* **Hung Hung** (National Taiwan University, Taiwan, ROC)
13:45-14:05	*Automatic Sparse Estimation of High-Dimensional Covariance Matrices* **Kazuyoshi Yata** (University of Tsukuba, Japan)
14:10-14:30	*Collaborative and Federated Black-box Optimization: A Bayesian Optimization Perspective* **Raed Al Kontar** (University of Michigan, USA)
14:35-14:55	*Effective Permutation Tests for Differences Across Multiple High-Dimensional Correlation Matrices* **Elio Zhang** (Seattle University, USA)

Parallel Sessions [Dec. 20, 13:20 – 15:00]

20c3 - Advances in Financial Econometrics and Network Modeling	
Location: S3	To Abstracts
Organizer: **Henghsiu Tsai** (Academia Sinica, Taiwan, ROC) **Mike K.P. So** (The Hong Kong University of Science and Technology, Hong Kong)	
Chair: **Cathy Chen** (Feng Chia University, Taiwan, ROC)	
13:20-13:45	*On Model Selection for Causal Inference* **Chor-yiu Sin** (National Tsing Hua University, Taiwan, ROC)
13:55-14:20	*Multi-View Dynamic Network Modeling* **Mike K.P. So** (The Hong Kong University of Science and Technology, Hong Kong)
14:30-14:55	*Probabilistic Loss Reserving Prediction via Denoising Diffusion Model* **Boris Choy** (The University of Sydney, Australia)

20c4 - Recent Advances in Data Science	
Location: S4	To Abstracts
Organizer: **Jian Huang** (The Hong Kong Polytechnic University, Hong Kong)	
Chair: **Hsuan-Yu Chen** (Academia Sinica, Taiwan, ROC)	
1	*Adaptive Debiased Lasso in High-Dimensional Generalized Linear Models with Streaming Data* **Yuanhang Luo** (The Hong Kong Polytechnic University, Hong Kong)
2	*DeepSuM: A Deep Sufficient and Efficient Modality Learning Framework* **Ting Li** (The Hong Kong Polytechnic University, Hong Kong)

Parallel Sessions [Dec. 20, 13:20 – 15:00]

20c5 - Lead Science and Clinical Research – Career Panel Discussion	
Location: S5	To Abstracts
Organizer: **Haoda Fu** (Amgen Inc., USA)	
Chair: **Haoda Fu** (Amgen Inc., USA)	
1	**Dacheng Liu** (Boehringer Ingelheim, Germany)
2	**Jingyuan Yang** (AbbVie, USA)
3	**Thomas Liu** (Amgen Inc., USA)
4	**Xun Chen** (AbbVie, USA)
5	**Lei Wang** (The Lotus Group, USA)

20c6 - Flexible Inference: Nonparametrics, Causality, and Large Language Models	
Location: A1	To Abstracts
Organizer: **Xiaowu Dai** (University of California, Los Angeles, USA)	
Chair: **Fang Liu** (University of Notre Dame, USA)	
13:20-13:45	*An Integrated GMM Shrinkage Approach with Consistent Moment Selection from Multiple External Sources* **Jun Shao** (University of Wisconsin, USA)
13:55-14:20	*Training-Free Multi-Agent Language Models* **Xiaowu Dai** (University of California, Los Angeles, USA)
14:30-14:55	*Optimal-PhiBE for Continuous-Time Reinforcement Learning with Discrete-Time Data* **Yuhua Zhu** (University of California, Los Angeles, USA)

Parallel Sessions [Dec. 20, 13:20 – 15:00]

20c7 - Recent Advances in Network Data Analysis	
Location: A2	To Abstracts
Organizer: **Haiyan Huang** (University of California, Berkeley, USA) **Yuguo Chen** (University of Illinois at Urbana–Champaign, USA)	
Chair: **Feng Liang** (University of Illinois at Urbana-Champaign, USA)	
13:20-13:45	*Statistically and Computationally Optimal Estimation and Inference in Common Subspaces* **Joshua Agterberg** (University of Illinois Urbana-Champaign, USA)
13:55-14:20	*UBSea: A Unified Community Detection Framework* **Hao Chen** (University of California, Davis, USA)
14:30-14:55	*Finding Anomalous Cliques in Inhomogenous Networks using Egonets* **Srijan Sengupta** (North Carolina State University, USA)

20c8 - Complex Data Analysis in Environmental and Health Studies	
Location: A3	To Abstracts
Organizer: **Zeny Feng** (University of Guelph, Canada)	
Chair: **Zeny Feng** (University of Guelph, Canada)	
13:20-13:40	*Tackling Explosive Likelihood and Non-Identifiability in Beta Mixtures* **Jiahua Chen** (The University of British Columbia, Canada)
13:45-14:05	*Robust Inverse Normal Transformation-Based Tests under Linear Mixed Effects Models* **Elif Acar** (University of Guelph, Canada)
14:10-14:30	*Using Statistical Modelling to Inform Public Health Decision Making in Hepatitis B* **William Wong** (University of Waterloo, Canada)
14:35-14:55	*Variable Selection in Mixture Regression Models* **Zeny Feng** (University of Guelph, Canada)

Parallel Sessions [Dec. 20, 13:20 – 15:00]

20c9 - Statistical Analyses Methods for Integrating Multi-Omics Data with Application to Personalized Medication	
Location: A4	To Abstracts
Organizer: **Maiying Kong** (University of Louisville, USA)	
Chair: **Maiying Kong** (University of Louisville, USA)	
1	*Multi-Dimensional Distributional Reinforcement Learning: A Hilbert Space Embedding Approach* **Qi Zheng** (University of Louisville, USA)
2	*Tuning Parameter Calibration for Prediction in Personalized Medicine* **Shih-Ting Huang** (University of Louisville, USA)

20c10 - Causal Inference for Complex Designs	
Location: A5	To Abstracts
Organizer: **Nancy R. Zhang** (University of Pennsylvania, USA) **Fan Li** (Duke University, USA)	
Chair: **Yen-Tsung Huang** (Academia Sinica, Taiwan, ROC)	
13:20-13:45	*Principal Stratification with U-Statistics under Principal Ignorability* **Fan Li** (Yale University, USA)
13:55-14:20	*Using a Two-Parameter Sensitivity Analysis Framework to Efficiently Combine Randomized and Non-randomized Studies* **Ruoqi Yu** (University of Illinois Urbana-Champaign, USA)
14:30-14:55	*Robust Sensitivity Analysis via Augmented Percentile Bootstrap under Simultaneous Violations of Unconfoundedness and Overlap* **Xinran Li** (University of Chicago, USA)

Parallel Sessions [Dec. 20, 15:20 – 17:00]:

20d1 - Causal Inference: Episode II	
Location: S1	To Abstracts
Organizer: **Jane-Ling Wang** (University of California, Davis, USA)	
Chair: **Fang Liu** (University of Notre Dame, USA)	
15:20-15:40	*Power and Sample Size Calculations for Causal Inference with Observational Data* **Fan Li** (Duke University, USA)
15:45-16:05	*Semiparametric mediation analysis using single-index models* **Yen-Tsung Huang** (Academia Sinica, Taiwan, ROC)
16:10-16:30	*Causal Mediation Analysis: A Summary-Data Mendelian Randomization Approach* **Shu-Chin Lin** (National Taiwan University, Taiwan, ROC)
16:35-16:55	*Sobolev Gradient Ascent for Optimal Transport: Barycenter Optimization and Convergence Analysis* **Changbo Zhu** (University of Notre Dame, USA)

20d2 - Frontiers in Knowledge Creation and Discovery	
Location: S2	To Abstracts
Organizer: **Masato Abe** (Doshisha University, Japan)	
Chair: **Chen-Hsiang Yeang** (Academia Sinica, Taiwan, ROC)	
15:20-15:40	*Integration of the Knowledge Space: Network Structure, Random Search and Synergy* **Masato S. Abe** (Doshisha University, Japan)
15:45-16:05	*Building A Food Knowledge Graph: Integration of Food-Related Data Sources* **Naoki Yoshimaru** (Doshisha University, Japan)
16:10-16:30	*Reproducibility-Optimized Component Clustering: A Test-Retest Framework for Robust Intrinsic Network Identification* **Arthur Chihhsin Tsai** (Academia Sinica, Taiwan, ROC)
16:35-16:55	*Function-on-Function Prediction via Deep Generative Models* **Tso-Jung Yen** (Academia Sinica, Taiwan, ROC)

Parallel Sessions [Dec. 20, 15:20 – 17:00]

20d3 - Advances in Stratified and Error-Prone Data Analysis	
Location: S3	To Abstracts
Organizer: **Peter Song** (University of Michigan, USA)	
Chair: **Ting Li** (The Hong Kong Polytechnic University, Hong Kong)	
15:20-15:45	*Novel Empirical Likelihood Method for the Cumulative Hazard Ratio under Stratified Cox Models* **Yichuan Zhao** (Georgia State University, USA)
15:55-16:20	*GUEST: Graphical Models for Ultrahigh-Dimensional and Error-Prone Data by the Boosting Algorithm* **Li-Pang Chen** (National Chengchi University, Taiwan, ROC)
16:30-16:55	*A Likelihood Approach for Data Integration Involving Missing Data and Misclassified Variables* **Hua Shen** (University of Calgary, Canada)

20d5 - Recent Advances in Statistical Learning for Complex Data Structures	
Location: S5	To Abstracts
Organizer: **Biao Cai** (City University of Hong Kong, Hong Kong) **Emma Jingfei Zhang** (Emory University, USA)	
Chair: **Hsueh-Han Huang** (Academia Sinica, Taiwan, ROC)	
15:20-15:40	*Trans-Ancestry Cell-Type-Specific eQTLs Mapping by Integrating scRNA-seq and Bulk Data* **Mingxuan Cai** (City University of Hong Kong, Hong Kong)
15:45-16:05	*Online Stochastic Optimization with Offline Data* **Yi Chen** (The Hong Kong University of Science and Technology, Hong Kong)
16:10-16:30	*Adaptive Bayesian Optimization with Consistent Smoothness Estimation and Hyperparameters Exploration* **Saifei Sun** (City University of Hong Kong, Hong Kong)
16:35-16:55	*A Preferential Latent Space Model for Text Networks* **Biao Cai** (City University of Hong Kong, Hong Kong)

Parallel Sessions [Dec. 20, 15:20 – 17:00]

20d6 - New Advances in Machine Learning and AI	
Location: A1	To Abstracts
Organizer: **Yi Li** (University of Michigan, USA)	
Chair: **Yan-Bin Chen** (National Taiwan University, Taiwan, ROC)	
15:20-15:45	*Penalized Generative Variable Selection* **Shuangge Ma** (Yale University, USA)
15:55-16:20	*Sparse Representation Learning for Scalable Single-Cell RNA Sequencing Data Analysis* **Zhixiang Lin** (The Chinese University of Hong Kong, Hong Kong)
16:30-16:55	(TBD) **Bin Nan** (University of California, Irvine, USA)

20d7 - Innovations in Machine Learning for Financial Data Analysis	
Location: A2	To Abstracts
Organizer: **Boris Choy** (The University of Sydney, Australia)	
Chair: **Nuttanan Wichitakorn** (Auckland University of Technology, New Zealand)	
15:20-15:40	*Bayesian Bi-directional Self-Exciting Threshold Autoregressive Models for Loss Reserving* **Wilson Chen** (The University of Sydney, Australia)
15:45-16:05	*Loss-based Bayesian Sequential Prediction of Value-at-Risk with a Long-Memory and Non-linear Realized Volatility Model* **Chao Wang** (The University of Sydney, Australia)
16:10-16:30	*Estimating Heterogeneous Treatment Effects through Multilevel Modeling* **Nuttanan Wichitakorn** (Auckland University of Technology, New Zealand)
16:35-16:55	*Predicting the Weekly Return Direction of the S&P 500 Index Using DNN for Time Series Classification* **Sang-Hyeok Lee** (Small Enterprise and Market Service, South Korea)

Parallel Sessions [Dec. 20, 15:20 – 17:00]

20d8 - Understanding the Biological Heterogeneity of Complex Traits Through Omics Data	
Location: A3	To Abstracts
Organizer: **Xiaofeng Zhu** (Case Western Reserve University, USA)	
Chair: **Hua Tang** (Stanford University, USA)	
15:20-15:40	*Investigating Spatial Omics Data with StarTrail and STimage-1K4M* **Yun Li** (University of North Carolina, USA)
15:45-16:05	*Multi-Ancestry Fine-Mapping of Causal Variants in Genome-Wide Association Studies* **Xiang Zhou** (Yale University, USA)
16:10-16:30	*An Alternative Method for Instrument Variable Regression: Reverse Two-Stage Least Squares (r2SLS)* **Wei Pan** (University of Minnesota, USA)
16:35-16:55	*Partitioned Blood Pressure Polygenic Risk Reveals Differential Genetic Effects and Environmental Modulation of Cardiovascular Disease* **Zhu Xiaofeng** (Case Western Reserve University, USA)

20d9 - Innovations in Survival Analysis and Clinical Trials for Biomedical Research	
Location: A4	To Abstracts
Organizer: **Yuanjia Wang** (Columbia University, USA)	
Chair: **Cheng-Shiun Leu** (Columbia University Irving Medical Center, USA)	
15:20-15:45	*Improving Unbiasedness of the Proportional Hazards Model Estimator with Cox and Snell's Bias Approximation and Jackknife Resampling* **Chia-Hui Huang** (National Chengchi University, Taiwan, ROC)
15:55-16:20	*Prediction-Oriented Transfer Learning for Semiparametric Transformation Models with Survival Data* **Yu Gu** (The University of Hong Kong, Hong Kong)
16:30-16:55	*A Novel Approach When Facing Emerging Infectious Disease --- Randomized Selection Design* **Cheng-Shiun Leu** (Columbia University Irving Medical Center, USA)

Abstract

December 17 (Wednesday)

December 18 (Thursday)

December 19 (Friday)

December 20 (Saturday)

Poster

December 17 (Wednesday):

Parallel Sessions [13:10 – 14:50]:

17a1 - Complex Analysis for Emerging Data
17a2 - Recent Developments on Generative AI and Statistics
17a3 - Innovations in Causal Mediation, Privacy, and AI-Driven Statistical Analysis
17a4 - Experimental Design
17a5 - Spatial Statistics
17a6 - Big Data and AI
17a7 - Quantum Computing in Statistics
17a8 - Innovative Bayesian Methods for Adaptive Clinical Trials
17a9 - Challenges in Survival Analysis and Clinical Trials
17a10 - Genomic and Multi-Omic

Parallel Sessions [15:10 – 16:50]:

17b1 - Memorial Session for Prof. Shaw-Hwa Lo
17b2 - Statistics and AI in Omics
17b3 - Spatial Statistics and Machine Learning
17b4 - Novel Methods for Event Prediction and Subgroup Identification
17b5 - Recent Advances in Lifetime Data Analysis
17b6 - Recent Research in Statistical Process Control, Part I
17b7 - Survival Analysis with Frailty and Copulas
17b8 - Recent Advances in Reliability Analysis
17b9 - Advanced Methods for Causal Inference
17b10 - Recent Advances in Bayesian Methods

Synthetic Data–Powered Statistical Inference with Generative AI

Xihong Lin

Department of Biostatistics and Department of Statistics, Harvard University

ABSTRACT

Scalable and robust statistical methods empowered by generative AI offer unprecedent potentials for trustworthy science as they empower statistical analysis, quantify uncertainty, enhance interpretability, and accelerate scientific discovery. In this talk, I will discuss robust and powerful statistical inference by leveraging synthetic data generated by generative AI models, such as diffusion models and transformer, while ensuring valid statistical inference even when generative AI models are misspecified. I will illustrate key points using the analysis of large scale biobanks, such as the analysis of the UK biobank whole genome and electronic health records, and demonstrate the power of scientific discovery by integrating statistics and generative AI using synthetic data.

Keywords: Generative AI; Synthetic data; Power; Statistical genetics; Biobanks

Additive Frechet Regression for Random Objects

Changwon Choi[1], Hans-Georg Mueller[2], **Byeong U. Park**[1], Wookyeong Song[2]

[1]*Department of Statistics, Seoul National University*

[2]*Department of Statistics, University of California, Davis*

ABSTRACT

Regression analysis for complex data taking values in a general metric space has gained increasing attention in recent years, particularly in the context of Frechet regression with Euclidean predictors X in R^p. However, nonparametric Frechet regression, while more flexible than global Frechet regression, suffers from the curse of dimensionality when the predictor dimension $p > 2$. To address this issue while maintaining modelling flexibility, we introduce a novel framework for additive structured nonparametric regression models with responses in general metric spaces. Due to the lack of vector space structure in general metric spaces where the responses reside, we propose a novel formulation that implicitly incorporates the additive structure via projection operators. Our method provides a unified framework for a wide range of response types, including distributions in Wasserstein space, network data represented by graph Laplacians, and spherical data equipped with geodesic distances. We establish consistency and derive convergence rates for the proposed additive \F regression estimators, leveraging smooth backfitting. The practical utility of our approach is demonstrated through applications to brain connectivity network analysis using resting-state fMRI data from Alzheimer's disease and cognitively normal subjects, as well as to distributional physical activity data from the NHANES study.

Keywords: Additive models; metric space valued responses; smooth backfitting; empirical process theory

Goodness-of-Fit and the Best Approximation: An Adversarial Approach

Qiwei Yao[1], Jinyaun Chang, Chengchun Shi, Mingcong Wu, Xinyang Yu

[1]*Department of Statistics, London School of Economics*

ABSTRACT

Diagnostic checking for goodness-of-fit is one of the important and routine steps in building a statistical model. The most frequently used approach for checking the goodness-of-fit is the residual analysis in the context of regression analysis. However for many statistical models there exist no natural residuals, which includes the models for the underlying distributions behind data, or the models for some complex dynamic structures such as the dynamic network models with dependent edges. Furthermore, there are scenarios in which there exist several competing models but none of them are the clear favourite. One then faces a task to choose the best approximation among the wrong models. We propose an adversarial approach in this paper. For checking the goodness-of-fit of a fitted model, we generate a synthetic sample from the fitted model and construct a classifier to classify the original sample and the synthetic sample into two different classes. If the fitted model is adequate, the classifier will have difficulties in distinguish the two samples. For identifying the best model among several candidate models, the classifier will create a distance between the original sample and the synthetic sample generated from each of the candidate model, and the model with the shortest distance is chosen as the best approximation for the truth.

Keywords: Classification, multi-layer perceptron, permutation test, sample-splitting

Assessing Spatial Spillover Effects in Mediation Analysis with Areal Data

Ritoban Kundu[1], **Peter X Song[2,*]**

Department of Biostatistics, University of Michigan

ABSTRACT

We consider a spatial mediation analysis that enables us to systematically capture the influence of neighboring regions on exposure-outcome relationships with areal data. We introduce a new framework of Linear Random Effects Spatial Structural Equation Model (LRES-SEM) that helps untangle the pathways under a directed acyclic graph (DAG) involving spatially distributed exposure, mediator and outcome. This new methodology allows to assess spatial spillover effects between spatially linked neighbors, extending the classical mediation analysis under the assumption of independent sampling units. Applying the LRES-SEM to analyze the county-level COVID-19 data across the U.S., we explore how political affiliation impacts mortality rates, both directly and through its influence mediated by vaccination hesitancy. Our results demonstrate the importance of accounting for the spatial connectivity, providing data evidence how political and social determinants may collectively impact health outcomes during a public health crisis like the COVID-19 pandemic.

Keywords: Causal pathway; Political ideology; Structural equation model; Vaccination hesitancy

Differentially Private Inference for Longitudinal Linear Regression

Getoar Sopa, **Marco Avella-Medina**, Cynthia Rush[1]
[1]*Columbia University, USA*

ABSTRACT

Differential Privacy (DP) provides a rigorous framework for releasing statistics while protecting individual information present in a dataset. Although substantial progress has been made on differentially private linear regression, existing methods almost exclusively address the item-level DP setting, where each user contributes a single observation. Many scientific and economic applications instead involve longitudinal or panel data, in which each user contributes multiple dependent observations; in these settings, item-level DP offers inadequate protection, and user-level DP - shielding an individual's entire trajectory - is the appropriate privacy notion. We develop a comprehensive framework for estimation and inference in longitudinal linear regression under user-level DP. We propose a user-level private regression estimator based on aggregating local regressions, and we establish finite-sample guarantees and asymptotic consistency under short-range dependence. For inference, we develop user-level private estimators of asymptotic covariance matrices via a privatized, bias-corrected covariance estimator that is automatically heteroskedasticity- and autocorrelation-consistent. These results provide the first unified framework for practical user-level DP estimation and inference in longitudinal linear regression under dependence, with strong theoretical guarantees and promising empirical performance.

Keywords: User-level differential privacy; Strong mixing; Concentration inequalities; Heteroscedasticity and autocorrelation consistent inference.

Learning from Shifted Data via a Semiparametric Selection Bias Model

Jiayang Sun[1], Zixiang Xu[1], Mary Meyer[2], Jing Qin[3]

[1]*Department of Statistics, George Mason University*

[2]*Colorado State University,* [3]*NIH*

ABSTRACT

Data shifts can occur in various forms, such as changes in covariates, label or prior distributions, domains, or data changes resulting from selection bias. This talk introduces some recent advances in treating data with potential selection bias and explores how these connect to broader data shift frameworks, including co-moving shifts. We present the estimation and testing procedures, demonstrating their effectiveness through theory, simulation studies, and applications to astronomy, and/or heart transplant data, time permitting.

Keywords: Data shifts, selection bias, isotonic inference, smoothing splines, heart transplant, SDSS

Orthogonalized Moment Aberration for Multi-Stratum Factorial Designs

Ming-Chung Chang

Institute of Statistical Science, Academia Sinica

ABSTRACT

Multi-stratum factorial designs, such as block designs and row–column designs, are widely used for screening treatment factors in experiments involving complex structures of experimental units due to multiple sources of error. In this study, we propose a unified model-free approach, termed orthogonalized moment aberration, to compare the similarities between level combinations of treatment factors assigned to heterogeneous experimental units. The proposed approach, which uses kernel functions to evaluate the rows of design matrices rather than the columns, can assess a wide variety of mixed-level regular/nonregular factorial designs with an extensive class of heterogeneous experimental unit structures called partially-relaxed orthogonal block structures. This approach is flexible in that it can be adapted to various scenarios by choosing different kernel functions, with certain choices yielding well-known minimum aberration criteria proposed in the literature. Although model-free, the proposed method is justified by using linear mixed- effect models and Gaussian process models. Theoretical results and numerical examples presented in this article collectively demonstrate that the proposed approach can generate multi-stratum factorial designs with high D-efficiencies within a Bayesian framework.

Keywords: Block design; Minimum aberration; Minimum moment aberration; Orthogonal array; Split-plot design

Efficient and Robust Block Designs for Order-of-Addition Experiments

Chang-Yun Lin

Department of Applied Mathematics and Institute of Statistics, National Chung Hsing University, Taichung, Taiwan, 40227

ABSTRACT

Designs for Order-of-Addition (OofA) experiments have received growing attention due to their impact on responses based on the sequence of component addition. In certain cases, these experiments involve heterogeneous groups of units, which necessitates the use of blocking to manage variation effects. Despite this, the exploration of block OofA designs remains limited in the literature. As experiments become increasingly complex, addressing this gap is essential to ensure that the designs accurately reflect the effects of the addition sequence and effectively handle the associated variability.

Motivated by this, this paper seeks to address the gap by expanding the indicator function framework for block OofA designs. We propose the use of the word length pattern as a criterion for selecting robust block OofA designs. To improve search efficiency and reduce computational demands, we develop algorithms that employ orthogonal Latin squares for design construction and selection, minimizing the need for exhaustive searches.

Our analysis, supported by correlation plots, reveals that the algorithms effectively manage confounding and aliasing between effects. Additionally, simulation studies indicate that designs based on our proposed criterion and algorithms achieve power and type I error rates comparable to those of full block OofA designs. This approach offers a practical and efficient method for constructing block OofA designs and may provide valuable insights for future research and applications.

Keywords: Component orthogonal array; indicator function; position-based; orthogonal Latin squares; word length pattern

Results on Large Strong Orthogonal Arrays of Strength Three

Chenlu Shi[1], Ye Tian[2], and Hongquan Xu[3,*]

[1]*Department of Mathematical Sciences, New Jersey Institute of Technology*

[2]*School of Mathematical Sciences, Beijing University of Posts and Telecommunications*

[3]*Department of Statistics and Data Science, University of California, Los Angeles*

ABSTRACT

Strong orthogonal arrays are widely recognized as effective space-filling designs for computer experiments. Among them, those of strength three are particularly useful, since strong orthogonal arrays of strength four or higher may be too expensive for some investigations. Strong orthogonal arrays of strength three that possess some of the space-filling properties of strength four are more desirable. Such arrays with small numbers of factors have been thoroughly investigated, whereas those of large sizes remain relatively unexplored. In this paper, we develop a characterization and construction method for large-sized arrays with better space-filling properties. We further present theoretical and computational results that facilitate the implementation of our construction method. Additionally, we use a simulation study to illustrate the usefulness of arrays produced by our method for developing statistical surrogate models.

Keywords: Computer experiment, Latin hypercube, space-filling design, strong orthogonal array

A Stratified L_2-Discrepancy with Application to Space-Filling Designs

Hongquan Xu*

Department of Statistics and Data Science, University of California, Los Angeles

ABSTRACT

Space-filling designs are widely used in computer experiments. We propose a stratified L_2-discrepancy to evaluate the uniformity of a design when the design domain is stratified into various subregions. Weights are used to adjust preferences for the uniformity over subregions in each stratification. The stratified L_2-discrepancy is easy to compute, satisfies a Koksma–Hlawka type inequality, and overcomes the curse of dimensionality that exists for other discrepancies. It is applicable to a broad class of designs, and covers several minimum aberration-type criteria as special cases. Strong orthogonal arrays of maximum strength are shown to have low stratified L_2-discrepancies, and thus are suitable for computer experiments. In addition, we develop a lower bound for the stratified L_2-discrepancy and provide a construction method for designs that achieve the lower bound. We further introduce a general version of the stratified L_2-discrepancy for evaluating designs with flexible stratification properties.

Keywords: Computer experiment; curse of dimensionality; generalised minimum aberration; space-filling hierarchy principle; strong orthogonal array.

Asymmetric Space–Time Covariance Functions via Hierarchical Mixtures

Pulong Ma

Department of Statistics, Iowa State University

ABSTRACT

This work is focused on constructing stationary space-time covariance functions through a hierarchical mixture approach that can serve as building blocks for capturing complex dependency structures. This hierarchical mixture approach provides a unified modelling framework that not only constructs a new class of asymmetric space-time covariance functions with closed-form expression, but also provides corresponding space-time process representations, which further unify constructions for many existing space-time covariance models. This hierarchical mixture framework decomposes the complexity of model specification at different levels of hierarchy, for which parsimonious covariance models can be specified with simple mixing measures to yield flexible properties and closed-form derivation. A characterization theorem is also provided for the hierarchical mixture approach on how the mixing measures determines the statistical properties of space-time covariance functions. A general class of asymmetric space-time covariance functions is also given that can allow arbitrary and possibly different degrees of smoothness in space and in time and flexible long-range dependence. Several of the new models are illustrated with simulation studies and a real dataset.

Keywords: Asymmetry; Hierarchical mixture; Smoothness; Long-range dependence; Space-time process

Graphical Modeling of Multivariate Inhomogeneous Spatial Point Processes

Junho Yang

Institute of Statistical Science, Academia Sinica

ABSTRACT

In this presentation, we propose a new graphical model for a multivariate spatial point process, where the intensity functions vary over space. The key idea is to utilize the coherence and partial coherence of the intensity-reweighted version of the original process. We introduce estimators for coherence and partial coherence and illustrate our results through a real data analysis of multivariate tropical forest tree data.

Keywords: Coherence; inhomogeneous; partial coherence; periodogram; point processes

Fast Variable Selection in Semiparametric Spatial Zero-inflated Models: Application to Extreme Climate Events

Chia-Ming Hsu, **Chun-Shu Chen**
Graduate Institute of Statistics, National Central University, Taiwan

ABSTRACT

Extreme climate events, such as heavy rainfall, are typically recorded as spatial count data with an overabundance of zeros due to their rarity. Modeling such zero-inflated spatial counts requires flexible approaches that simultaneously account for spatial dependence and excess zeros. In this talk, we present a semiparametric spatial zero-inflated modeling framework tailored to the analysis of extreme climate events. The proposed method introduces a novel, distribution-free, and computationally efficient variable selection criterion, inspired by the structure of Lasso regression, to identify key covariates without reliance on a fully specified likelihood. We demonstrate its performance through comprehensive simulations. An application to Taiwan's 2016 daily extreme rainfall data highlights the method's practical utility for environmental studies.

Keywords: Computational efficiency; Extreme events; Lasso regression; Model selection; Spatial count data

Leveraging AI techniques to Study the Risk of Accelerated Brain Aging and Dementia Using Large-Scale Biobank Data

Tianzhou Ma[1], Menglu Liang[1], Zhenyao Ye[2], Neng Wang[1], Li Feng[1], Shuo Chen[2]

[1]*University of Maryland College Park*

[2]*University of Maryland School of Medicine*

ABSTRACT

As we age, brain structure and function decline - reflected in reduced volume, cortical thinning, white-matter deterioration and altered functional connectivity - leading to cognitive decline and elevated dementia risk. However, there is significant individual variation in how fast the brain ages and the timing of when these changes emerge and progress into age-related diseases. Large biobank data like UK Biobank (UKB) have made vast amounts of individual genetic, imaging, electronic health record and lifestyle/health risk data accessible and created an unprecedented opportunity for brain aging and dementia research. We applied machine learning method to predict white matter brain age from regional diffusion tensor imaging data in UKB and calculate Brain Age Gap (BAG), which serves as a promising biomarker for early dementia detection. We then conducted a number of Mendelian Randomization studies to identify the nongenetic risk factors (including smoking, blood pressure, allostatic load and other lifestyle factors) of accelerated brain aging. Considering the complex polygenic nature of brain aging, we developed a novel pathway-guided multivariate transcriptome-wide association studies (TWAS) method and embedded sparse group lasso within the method to select genes and pathways most associated with brain aging, improving our understanding of the molecular mechanism of the aging brain and the transition to dementia. Existing dementia risk prediction models were developed almost exclusively in White populations of European ancestry and perform poorly in racial and ethnic minority groups, perpetuating health disparities in dementia prevention and treatment. Here we developed a deep transfer learning model to predict dementia risk in racial minority populations. We included ~300 candidate risk factors for ~500,000 participants from White, Black and Asian populations in UKB to train and test the model, and used the health data of ~300,000 participants from US-based All of Us cohort for external validation. Our approach achieved superior predictive performance for minority populations than existing risk prediction models while identifying both common and population-specific risk factors to construct interpretable population-specific dementia risk scores for clinical usage.

Keywords: white matter, brain aging, dementia, sparse group lasso, transfer learning, health disparity

On MCMC Mixing for Predictive Inference under Unidentified Transformation Models

Chong ZHONG[2*], Jin YANG[2*], Junshan SHEN[2*], Zhaohai Li[2*], **Catherine C. Liu[1]**

Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University

ABSTRACT

Reliable Bayesian predictive inference has long been an open problem under unidentified transformation models, since the Markov Chain Monte Carlo (MCMC) chains of posterior predictive distribution (PPD) values are generally poorly mixed. In this talk, we introduce how we address the poorly mixed PPD value chains under unidentified transformation models through an adaptive scheme for prior adjustment.

Specifically, we originate a conception of sufficient informativeness, which explicitly quantifies the information level provided by nonparametric priors, and assesses MCMC mixing by comparison with the within-chain MCMC variance. We formulate the prior information level by a set of hyperparameters induced from the nonparametric prior elicitation with an analytic expression, which is guaranteed by asymptotic theory for the posterior variance under unidentified transformation models. The analytic prior information level consequently drives a hyperparameter tuning procedure to achieve MCMC mixing. The proposed method is general enough to cover various data domains through a multiplicative error working model. Comprehensive simulations and real-world data analysis demonstrate that our method successfully achieves MCMC mixing and outperforms state-of-the-art competitors in predictive capability.

Keywords: Bayesian nonparametrics; MCMC mixing; Predictive inference; Prior information level

Microbiome Data Integration via Shared Dictionary Learning

Bo Yuan[1] and **Shulei Wang[1]**

Department of Statistics, University of Illinois Urbana-Champaign

ABSTRACT

Data integration is a powerful tool for facilitating a comprehensive and generalizable understanding of microbial communities and their association with outcomes of interest. However, integrating data sets from different studies remains a challenging problem because of severe batch effects, unobserved confounding variables, and high heterogeneity across data sets. We propose a new data integration method called MetaDICT, which initially estimates the batch effects by weighting methods in causal inference literature and then refines the estimation via novel shared dictionary learning. Compared with existing methods, MetaDICT can better avoid the overcorrection of batch effects and preserve biological variation when there exist unobserved confounding variables, data sets are highly heterogeneous across studies, or the batch is completely confounded with some covariates. Furthermore, MetaDICT can generate comparable embedding at both taxa and sample levels that can be used to unravel the hidden structure of the integrated data and improve the integrative analysis. Applications to synthetic and real microbiome data sets demonstrate the robustness and effectiveness of MetaDICT in integrative analysis. Using MetaDICT, we characterize microbial interaction, identify generalizable microbial signatures, and enhance the accuracy of outcome prediction in two real integrative studies, including an integrative analysis of colorectal cancer metagenomics studies and a meta-analysis of immunotherapy microbiome studies.

Keywords: Microbiome Data Integration, Batch Effect, Dictionary Learning

Quantum Speedups for Multiproposal MCMC

Chin-Yi Lin[1], Kuo-Chin Chen[1*], Philippe Lemey[2], Marc A. Suchard[3], Andrew J. Holbrook[4], Min-Hsiu Hsieh[1**]

[1]*Foxconn Research*

[2]*Department of Microbiology, Immunology and Transplantation, KU Leuven*

[3]*Departments of Biostatistics, Biomathematics and Human Genetics, UCLA, Los Angeles, USA*

[4]*Department of Biostatistics, UCLA*

ABSTRACT

Multiproposal Markov chain Monte Carlo (MCMC) algorithms choose from multiple proposals to generate their next chain step in order to sample from challenging target distributions more efficiently. However, on classical machines, these algorithms require O(P) target evaluations for each Markov chain step when choosing from P proposals. Recent work demonstrates the possibility of quadratic quantum speedups for one such multiproposal MCMC algorithm. After generating P proposals, this quantum parallel MCMC (QPMCMC) algorithm requires only O(P) target evaluations at each step, outperforming its classical counterpart. However, generating P proposals using classical computers still requires time complexity O(P), resulting in the overall complexity of QPMCMC remaining O(P). Here, we present a new, faster quantum multiproposal MCMC strategy, QPMCMC2. With a specially designed Tjelmeland distribution that generates proposals close to the input state, QPMCMC2 requires only O(1) target evaluations and Olog(P) qubits when computing over a large number of proposals P. Unlike its slower predecessor, the QPMCMC2 Markov kernel (1) maintains detailed balance exactly and (2) is fully explicit for a large class of graphical models. We demonstrate this flexibility by applying QPMCMC2 to novel Ising-type models built on bacterial evolutionary networks and obtain significant speedups for Bayesian ancestral trait reconstruction for 248 observed salmonella bacteria.

Keywords: Bayesian phylogenetics; MCMC; Quantum algorithms; Ising models

A Derivative-Free Approach for Parameter Inference in Hidden Quantum Markov Models

Ning Ning

Department of Statistics, Texas A&M University, College Station, USA

ABSTRACT

Hidden Quantum Markov Models (HQMMs) provide a quantum-inspired framework for modeling complex sequential data, offering greater expressive power than classical Hidden Markov Models (HMMs). Existing learning algorithms for HQMMs typically rely on gradient-based optimization of the log-likelihood function, which can be computationally intensive and sensitive to local minima.

In this work, we propose a new and general method for inferring HQMM parameters that avoids the computation of derivatives of the log-likelihood. Our approach is broadly applicable to HQMMs with arbitrary Kraus operator structures and enables learning in settings where differentiability is difficult to guarantee or gradient computation is costly. We validate the proposed algorithm on synthetic datasets, demonstrating that it can successfully recover HQMM parameters and outperforms baseline methods in both accuracy and computational efficiency.

Keywords: Quantum computing; Hidden Markov Models; Sequential Data Modeling; Parameter Inference; Derivative-Free Optimization

Quantum Computations of Partial Differential Equations and Related Problems

Shi Jin

Institute of Natural Sciences, Shanghai Jiao Tong University

ABSTRACT

Quantum computers are designed based on quantum mechanics principle, they are most suitable to solve the Schrodinger equation, and linear PDEs (and ODEs) evolved by unitary operators. It is important to explore whether other problems in scientific computing, such as ODEs, PDEs, and linear algebra that arise in both classical and quantum systems which are not unitary evolution, can be handled by quantum computers.

We will present a systematic way to develop quantum simulation algorithms for general differential equations. Our basic framework is dimension lifting, that transfers non-autonomous ODEs/PDEs systems to autonomous ones, nonlinear PDEs to linear ones, and linear ones to Schrodinger type PDEs—coined "Schrodingerization"—with unitary evolutions. Our formulation allows both qubit and qumode (continuous-variable) formulations, and their hybridizations, and provides the foundation for analog quantum computing which are easier to realize in the near term. We will also present dimension lifting techniques for quantum simulation of stochastic DEs and PDEs with fractional derivatives, and quantum machine learning. A quantum simulation software—"UnitaryLab"—will also be introduced.

Keywords: quantum algorithms, Schrodingerization, partial differential equations, analog quantum computing, dimension lifting

Adaptive Circuit Learning of Born Machine: Towards Realization of Amplitude Embedding and Quantum Data Loading

Chun-Tse Li, **Hao-Chung Cheng***

Department of Electrical Engineering, National Taiwan University, Taipei 10639, Taiwan

Hon Hai Quantum Research Center, Taipei 114, Taiwan

ABSTRACT

Quantum data loading plays a central role in quantum algorithms and quantum information processing. Many quantum algorithms hinge on the ability to prepare arbitrary superposition states as a subroutine, with claims of exponential speedups often predicated on access to an efficient data-loading oracle. In practice, constructing a circuit to prepare a generic n-qubit quantum state typically demands computational efforts scaling as $O(2^n)$, posing a significant challenge for quantum algorithms to outperform their classical counterparts. To address this critical issue, various hybrid quantum–classical approaches have been proposed. However, many of these solutions favor simplistic circuit architectures, which are susceptible to substantial optimization challenges. In this study, we harness quantum circuits as Born machines to generate probability distributions. Drawing inspiration from methods used to investigate electronic structures in quantum chemistry and condensed matter physics, we propose a framework called Adaptive Circuit Learning of Born Machine, which dynamically expands the ansatz circuit. Our algorithm is designed to selectively integrate two-qubit entangled gates that best capture the intricate entanglement present within the target state. Empirical experiments underscore the efficacy of our approach in encoding real-world data through amplitude embedding, demonstrating not only compliance with but also enhancement over the performance benchmarks set by prior research.

Keywords: quantum data loading; born machine; quantum machine learning; amplitude encoding

Randomized Optimal Selection Design for Dose Optimization

Shuqi Wang, **Ying Yuan**, and Suyu Liu

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, U.S.A.

ABSTRACT

The U.S. Food and Drug Administration (FDA) launched Project Optimus to shift the objective of dose selection from the maximum tolerated dose to the optimal biological dose (OBD), optimizing the benefit-risk tradeoff. One approach recommended by the FDA's guidance is to conduct randomized trials comparing multiple doses. In this paper, using the selection design framework, we propose a randomized optimal selection (ROSE) design, which minimizes sample size while ensuring the probability of correct selection of the OBD at prespecified accuracy levels. The ROSE design is simple to implement, involving a straightforward comparison of the difference in response rates between two dose arms against a predetermined decision boundary. We further consider a two-stage ROSE design that allows for early selection of the OBD at the interim when there is sufficient evidence, further reducing the sample size. Simulation studies demonstrate that the ROSE design exhibits desirable operating characteristics in correctly identifying the OBD. A sample size of 15 to 40 patients per dosage arm typically results in a percentage of correct selection of the optimal dose ranging from 60% to 70%. A user-friendly software for implementing ROSE designs is available on www.trialdesign.org.

Keywords: Dose optimization; Optimal design; Randomization; Selection design.

Exploring Sensitive Biomarkers with Short-Term Response and Long-Term Outcome Using Bayesian Additive Regression Trees

Satoshi Morita, Zixuan Yao

Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine, Kyoto, Japan

e-mail: smorita@kuhp.kyoto-u.ac.jp

ABSTRACT

Identifying good predictive baseline covariates for optimizing the target population for a new treatment is a topic that has attracted great interest. In some situations, an early post-baseline biomarker response may serve as a supportive guide for physicians to decide whether to continue a treatment for a patient. We propose an exploratory two-stage subgroup-analysis method as a statistical tool to investigate a role of such a short-term outcome in informing each individual patient of whether they benefit from a new treatment in terms of long-term outcome. We use a flexible probability model, Bayesian additive regression trees (BART), to derive predictive conditional treatment effects (PCTE) in a short-term post-baseline biomarker response based on counter-factual modeling of responses to new and standard treatments for each patient. Constructing patient subgroups according to the PCTE values, we analyze an observed long-term outcome to explore a sensitive subpopulation. We carry out extensive simulation studies to examine the operating characteristics of the proposed method. For illustration, we apply the proposed method to data from a randomized clinical trial in oncology.

Keywords: Bayesian additive regression trees; Short-term biomarker response; Time-to-event; Sensitive subpopulation; Randomized clinical trials.

BIT: A Bayesian Optimal Adaptive Clinical Trial Design for Integrated Therapies

Yong Zang

Department of Biostatistics and Health Data Sciences,

Indiana University School of Medicine

ABSTRACT

Complex chronic diseases often require integrated therapies administered sequentially across different disease phases. In alcohol-associated hepatitis (AH) and alcohol use disorder (AUD), optimizing treatment selection is essential for improving long-term survival. In this talk, we introduce a Bayesian adaptive design specifically tailored for integrated therapy trials (BIT). The BIT design employs flexible Bayesian parametric modelling approaches to characterize therapeutic effects across disease phases and incorporates multiple interim analyses with adaptive stopping rules for both futility and superiority to enhance efficiency while strictly controlling the family-wise type I error rate and maximizing statistical power. Simulation studies confirm the desirable operating characteristics of the BIT design. While motivated by AH/AUD, the proposed framework is broadly applicable to other complex chronic diseases requiring sequential treatment strategies.

Keywords: Adaptive design, Bayesian statistics, Clinical trial

An Empirical Bayesian Method for Subgroup Identification in Personalized Medicine

Jian Yin[1], Zihang Zhong[1], Senmiao Ni[1], Yang Zhao[1], **Jingwei Wu[2]**, Hao Yu[1], Jianling Bai[1]

[1]*Department of Biostatistics, School of Public Health, Nanjing Medical University*

[2]*Department of Epidemiology and Biostatistics, Barnett College of Public Health, Temple University*

ABSTRACT

Personalized medicine tailors therapies to patient-specific characteristics, creating a growing need for reliable subgroup identification. Data-driven approaches that search for subgroups with enhanced efficacy or safety using predictive biomarkers are especially valuable when mechanistic knowledge is limited. However, single-trial methods often suffer from selection inaccuracies and instability due to insufficient information. To address these limitations, we propose PEMBA (normalized Power prior based on the EMpirical BAyesian method), a subgroup identification framework that improves detection of treatment-effect heterogeneity by incorporating evidence from multiple historical studies together with current trial data. PEMBA applies a normalized power prior within an empirical Bayesian structure to integrate information and compute posterior treatment-effect distributions through a grid search for optimal splits. A permutation test is used to control the overall false positive rate. Simulation studies show that PEMBA improves subgroup classification accuracy compared with existing approaches while maintaining false positive rates at pre-specified levels. The method also remains robust in the presence of heterogeneity across historical trials. A real-data application in breast cancer further demonstrates PEMBA's ability to leverage multi-trial information to identify a clinically meaningful subgroup. By integrating evidence across studies, PEMBA provides a more reliable approach to detecting treatment-effect heterogeneity. This method can advance personalized medicine by improving clinical study efficiency, increasing trial success rates, and enabling more patients to receive appropriately targeted treatments.

Keywords: Subgroup Identification, Data borrowing, Normalized power prior

Dynamic and Concordance-Assisted Learning for Risk Stratification

Jing Ning

Department of Biostatistics/1MC 12.3557

The University of Texas M.D. Anderson Cancer Center

7007 Bertner Ave Houston, TX 77030 USA

ABSTRACT

Dynamic prediction models that adapt over time and maintain accuracy can play a crucial role in monitoring disease progression in clinical practice. In biomedical studies with long-term follow-up, participants are typically monitored through periodic clinical visits with repeated measurements until the occurrence of the event of interest (e.g., disease onset) or study completion. Recognizing the dynamic nature of disease risk and the information captured by longitudinal markers, we propose an innovative concordance-assisted learning algorithm to derive a real-time risk stratification score. Our approach avoids the need to fit regression models, such as joint models of longitudinal markers and time-to-event outcomes, thereby offering robustness to model misspecification. Simulation studies demonstrate that the proposed method performs well in dynamically monitoring disease risk and distinguishing high- from low-risk populations over time. We further apply the method to data from the Alzheimer's Disease Neuroimaging Initiative to develop a dynamic risk score for Alzheimer's disease in patients with mild cognitive impairment, incorporating multiple longitudinal markers and baseline prognostic factors.

Keywords: concordance-assisted learning; dynamic prediction; longitudinal markers; risk stratification.

Model Based Multiple Imputation in Censored Quantile Regression

Zhaozhi Fan, Ummay Nayeema Islam

Department of Mathematics and Statistics, Memorial University of Newfoundland

ABSTRACT

Censored quantile regression models provide a global description of the association between the censored response and potential risk factors, through proper selection of quantiles. But when the censoring rate goes high, the model would face identifiability issues, especially for the extreme quantiles. In this article we propose a multiple imputation method based on the AFT model to handle the censored values. Quantile regression parameters are estimated based on the imputed data. The imputation was done multiple times through randomly selecting residuals from AFT regression. The estimators of the regression parameters are consistent and having asymptotic multivariate normal distribution. Simulation results show that our method performs equally well with existing approaches when the models are identifiable and can also provide reliable estimation beyond the identifiable range of quantile levels of existing methods.

Keywords: survival analysis; quantile regression; multiple imputation

Improving Single-Cell Perturbation Analyses through Efficiency Estimation

Qiyuan Liu[1], Siming Zhao[2], **Jingshu Wang[1]**

[1] *Department of Statistics, The University of Chicago*

[2] *Department of biomedical data science and the Cancer Center, Dartmouth College*

ABSTRACT

Single-cell perturbation screening has transformed functional genomics, yet its utility is hampered by noisy data and low statistical power. A central but often overlooked challenge is perturbation efficiency: some perturbations strongly alter target gene expression, while others fail to induce measurable changes, even within the same experiment. Ignoring this heterogeneity renders causal estimands dataset-dependent and undermines reproducibility.

We show that existing approaches for estimating cell-level perturbation efficiency are vulnerable to confounding, leading to biased effect estimates and false discoveries. To address this, we develop an instrumental-variable framework that treats perturbations as instruments rather than direct treatments. By estimating gRNA-specific efficiencies—shifting resolution from individual cells to groups of cells—we obtain consistent estimates that reveal striking variability across gRNAs targeting the same gene. Incorporating these efficiencies into differential expression analyses improves detection power, yields more interpretable causal effect sizes, and enhances consistency across datasets. When sample sizes permit, our framework further enables investigation of dosage effects, providing a clearer picture of heterogeneity in gene perturbation responses.

Our results highlight the importance of rigorously defining and estimating perturbation efficiency, thereby improving both the validity and interpretability of single-cell perturbation studies.

Keywords: instrumental variable, single-cell genomics, CRISPR screening, causal effect estimation

Addressing heterogeneous sensitivity in biomarker screening with application in NanoString nCounter data

Chang Yu, **Zhijin Wu**

Department of Biostatistics, Brown University

ABSTRACT

Biomarkers are measurable indicators of biological processes and have wide biomedical applications including disease screening and prognosis prediction. Candidate biomarkers can be screened in high-throughput settings, which allow simultaneous measurements of a large number of molecules. The ability to detect a molecule may be hindered by the presence of background noise and the variable signal strength, which depends on both the properties of the target molecule and the quality of the sample. The detection sensitivity thus varies in a marker- and sample-specific manner. This heterogeneity in detection sensitivity is often overlooked and leads to an inflated false positive rate. We propose a novel *sensitivity adjusted likelihood-ratio test* (SALT), which properly controls the false positives and is more powerful than the unadjusted approach. We show that sample-and-feature-specific detection sensitivity can be well estimated from NanoString nCounter data, and using the estimated sensitivity in SALT results in improved biomarker screening.

Keywords: High-throughput screening, Biomarker, Sensitivity, NanoString nCounter

The Taiwan Precision Medicine Initiative: Building the Largest non-European Cohort for Precision Health

Jer-Yuarn Wu

Institute of Biomedical Sciences, Academia Sinica

ABSTRACT

Data from the national biobanks, especially those from the UK Biobank, show that disease risk prediction based on genetic profiles is feasible but the prediction models created with European data do not translate to other ethnic groups. The current bottleneck is therefore a large dataset which consists of genetic profiles and matching clinical information from many individuals in Taiwan. Equally important is the analysis of this large dataset to produce disease risk models that predict an individual's disease risk for all the major diseases. To establish this large dataset, we collaborated with 16 major medical systems across Taiwan and launched Taiwan Precision Medicine Initiative (TPMI) in 2019.

We produced a reference panel against which anyone of Han Chinese ancestry can use to assess his/her own disease risk for precision health management. Leveraging the whole genome sequencing data produced by the Taiwan Biobank and the genome reference assemblies produced by the Kwok's group, we have designed and validated custom single nucleotide polymorphism (SNP) arrays that can produce the genetic profile of a person and test all known clinically useful genetic variants at the same time. We have genotyped ~500,000 participants enrolled (as of December 31, 2023). All participants have agreed to contribute their clinical data (EMR, electronic medical record) to the database, allowing us to analyze not just their current disease state, but also their treatment response and long term disease progression.

Our result showed that 327 diseases have more than 10,000 participants. 37 of those diseases have polygenic risk scores with high accuracy (AUC value larger than 0.6). We anticipate the established TPMI dataset will advance the promise of "Precision Medicine" for individuals in Taiwan.

Key words: polygenic risk score, EMR (electronic medical record), AUC

A Multi-Tissue Map of Protein Regulation Reveals Shared and Context-Dependent Genetic Architectures

Hua Tang[1], Huaying Fang[1,2], Lihua Jiang[1]

[1]*Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA.*

[2]*Academy for Multidisciplinary Studies, Capital Normal University, Beijing 100048, China.*

ABSTRACT

The abundance of proteins, essential functional units of the cell, is tightly controlled by genetic and environmental factors. Dysregulation of proteins is fundamental to disease. While transcriptomic and plasma proteomic studies have provided key insights into molecular regulation, a tissue-resolved regulatory map across normal human organs remains elusive. Here we quantify over 10,000 unique proteins across five disease-relevant tissues from hundreds of donors, establishing a systematic tissue proteome map. We identify nearly 2,000 cis-regulatory loci, most not previously linked to protein abundance in plasma, and show that these local effects are broadly shared across tissues. In contrast, trans-acting loci and sex associations are highly context-specific, while age associations show intermediate sharing. Integration with transcriptomics further reveals that, within a tissue, gene-level RNA–protein correlations across individuals are generally low, underscoring the pervasive role of post-transcriptional regulation. Together with existing plasma studies, these findings define a layered architecture of protein regulation that spans tissues and plasma, illuminating both shared and tissue-specific biology. This resource provides a molecular framework for connecting genetic and biological factors to protein regulation and for advancing mechanistic insight into human complex diseases.

Keywords: proteomics, multi-omics, normal human tissues, pQTL, gene-regulation

Bayesian Rhythmic Model for Jointly Detecting Circadian Biomarkers and Predicting Molecular Circadian Time in Human Post-Mortem Brain Transcriptome

Xiangning Xue[1], Liangliang Zhang[2], **George Tseng[1]**

[1]*Department of Biostatistics and Health Data Science, University of Pittsburgh*

[2]*Department of Population and Quantitative Health Sciences, Case Western Reserve University.*

ABSTRACT

Transcriptomic circadian analysis in human post-mortem brain has provided an unprecedented opportunity to decipher in vivo zzhowmolecular circadian rhythms in different human brain regions that are known to play essential roles in aging and psychiatric disorders. Detecting circadian biomarkers is often the first objective in the statistical analysis. Confounded with the task is the common issue that the true molecular circadian clock time within a human subject is usually inconsistent with the record, which could come from observation or recording errors, or the intrinsic biological variation. In the literature, many methods have been proposed for detecting rhythmic biomarkers or for predicting true molecular circadian time, respectively. To date, no method is developed to achieve both objectives simultaneously. In this paper, we propose a Bayesian model for simultaneous Circadian marker detection and molecular circadian Time estimation (BayCT). The model is further extended to repeated measure of multiple tissues or brain regions. We adopt Von Moses prior distribution for angular data with slice sampling and reversible jump sampling in the Markov chain Monte Carlo (MCMC) procedure for Bayesian inference. We demonstrate the method by extensive simulations and two applications in 12 tissues in mouse brain and three brain regions in human. The result shows superior performance in both statistical objectives with large margin in both circadian marker detection and circadian time detection, as well as the advantage of integrating multiple brain regions.

Keywords: circadian rhythm, omics analysis, Bayesian model

Adjusting Transcript Leakage in Spatial Transcriptomic Data

Christina Huan Shi[1,#], Yibo Zhai[2,#], Savio Ho-Chit Chow[1], Liangbang Li[2], Peter D. Adams[1], Bing Ren[3], Marissa Schafer[4], **Yingying Wei[2,*]**, Kevin Y. Yip[1,*]

[1]*Sanford Burnham Prebys Medical Discovery Institute.*

[2]*Department of Statistics and Data Science, The Chinese University of Hong Kong*

[3]*Department of Systems Biology, Columbia University*

[4]*Mayo Clinic*

[#]*Contributed equally*

ABSTRACT

Spatial transcriptomics offers unprecedented opportunities to investigate the tissue organization, cell-cell interactions and the tumour microenvironment. Compared with spot-based spatial transcriptomic technologies, which typically assay multiple cells within a single regular-shaped spot and hence yield bulk-level data, recent spatial transcriptomic platforms can achieve cellular or even subcellular resolution. However, recent benchmarking studies of spatial transcriptomic platforms have reported that certain genes appear to be expressed in cell types where such expression is biologically unexpected, even after applying highly stringent cell segmentation and filtering procedures. Our data analysis suggests that this phenomenon is likely attributable to transcript leakage from neighboring cells. However, despite active research on the deconvolution of spot-level spatial transcriptomic data, statistical methods for decontaminating spatial transcriptomic data affected by transcript leakage are lacking. Unfortunately, if the transcript leakage problem is not properly adjusted, it can lead to serious consequences for the quantification of expression levels, cell type annotation, differential gene expression detection and identification of spatially variable genes. Here, we have developed a Bayesian hierarchical model to adjust gene expression contamination resulting from transcript leakage. We have proven the model identifiability, which shows that contamination due to transcript leakage can be distinguished from the true underlying gene expression. Application to real data shows that our model successfully adjusts transcript leakage in spatial transcriptomic data.

Keywords: Spatial Transcriptomics; Deconvolution; Identifiability; Bayesian Hierarchical Model.

Computational Intelligence from Omics to Medicine

Ka-Chun Wong

Department of Computer Science, City University of Hong Kong

ABSTRACT

In recent years, the integration of Artificial Intelligence (AI) in scientific research has revolutionized the field of molecular biology and medicine. This speech aims to explore three significant aspects of computational intelligence in molecular biology and medicine. In particular, I will navigate my research group efforts from omics to medicine: bioinformatics, medical informatics, and clinical informatics. By leveraging computational intelligence, my research group has enabled multiple advances in DNA motif analysis, cancer detection, gene editing, and small-molecule drug discovery.

1. In bioinformatics, pattern recognition algorithms on DNA motifs are presented and demonstrated to be instrumental in identifying and understanding the intricate patterns within DNA sequences. These algorithms aid in deciphering gene regulatory elements, enabling researchers to unravel the complexity of genetic networks and their impact on cellular processes, enabling insights into the fundamental mechanisms governing gene expression and regulation.

2. In medical informatics, machine learning algorithms are presented to demonstrate accuracy in analysing complex medical data. By training models on vast datasets, the proposed algorithms can identify subtle patterns indicative of cancerous cells, assisting in early detection and precise localization of tumours, leading to improved patient outcomes and personalized treatment strategies.

3. In clinical informatics, several computational intelligence approaches are presented for gene editing and small-molecule drug discovery. This enables precise gene editing, offering potential therapeutic interventions for genetic disorders. Last but not least, drug docking techniques are proposed to accelerate the identification of small-molecule compounds with diffusion modelling.

Keywords: AI for Science; Bioinformatics; Medical Informatics; Gene Editing; Drug Discovery

Estimation and Selection in Survival Models for Individuals with Spatial Frailty

Kyeongeun Kim[1], Joonho Shin[2], **Chae Young Lim[1]**

[1]*Department of Statistics, Seoul National University, Korea*

[2]*School of Mathematics, Statistics and Data Science, Sungshin Women's University, Korea*

ABSTRACT

In this talk, we introduce an estimating equation approach for a right-censored survival model with spatial random effects called spatial frailty. Our method accommodates multiple individuals at each site and incorporates spatial dependence of the individuals across sites. Additionally, we consider penalization for group variable selection, enabling the identification of key group variables, including categorical covariates, that influence survival times. We establish consistency and asymptotic normality as well as the oracle property of our estimator. A simulation study with various scenarios supports our theoretical findings. We also apply the proposed approach to real data, analyzing the survival of businesses in developing commercial districts of Seoul, South Korea, with a focus on spatial dependency using location-based map data. Our results suggest that the proposed method is a useful tool for analyzing spatially correlated survival data.

Keywords: Estimating equation; Spatial frailty; Spatial survival analysis; Variable selection

A Spatio-Temporal Modeling Approach for Wind Speed Data from a Regional Climate Model

Eva Murphy[1], Whitney Huang[2], Ting Fung Ma[3]

[1]Department of Statistical Sciences, Wake Forest University

[2]School of Mathematical and Statistical Sciences, Clemson University

[3]Department of Statistics, University of South Carolina

ABSTRACT

Statistical models that describe how physical processes vary across space and time are essential in environmental studies. We propose a spatio-temporal modeling framework for wind speed that is continuous in space and discrete at regular time intervals. The model adopts an additive decomposition: a smooth space–time function captures the mean structure, incorporating temporal periodicity associated with the annual cycle, while a combination of empirical orthogonal functions (EOFs) and a first-order dynamical Gaussian process enables a nonstationary spatial covariance structure and a parsimonious temporal dependence. The proposed approach is applied to regional climate model data from Canada, demonstrating improved and more consistent predictive performance compared to a baseline method. This is joint work with Eva Murphy at Wake Forest University and Ting Fung Ma from University of South Carolina.

Keywords: Spatio-temporal modeling; Empirical orthogonal functions; Gaussian processes; Wind Speeds

A Comparative Study of Neural Network Adaptations for Spatial Data

Bo-yu Chen[1] and Hao Zhang[2]

[1]*Department of Statistics, Purdue University*

[2]*Department of Statistics and Probability, Michigan State University*

ABSTRACT

In this study, we evaluated the predictive performance of various deep neural network approaches for spatial data. Recent advances in neural networks have led to a surge in deep learning methods and applications. Many of these methods were developed implicitly for independent data, and it is not trivial to incorporate spatial correlation into them. However, several strategies have been proposed to adapt neural networks to account for spatial correlation. Using both simulated and real-world datasets, we compared spatial methods developed from fully connected deep neural networks, analyzing the impact of different spatial adaptations. Our findings aim to inform future research directions in this evolving field.

Keywords: Deep learning, Kriging, Neural network, Prediction, Spatial statistics

Spectral Radii of Kernel Matrices and Applications to Kernel Score Tests

Hao Zhang

Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA

ABSTRACT

In spatial models and linear mixed-effects models, a central question is whether the data exhibit independence or, alternatively, whether spatial or random effects are present. One approach, derived from the score function (the derivative of the log-likelihood), was first developed in the context of linear mixed-effects models using the dot-product kernel. This framework was later generalized to incorporate the Gaussian kernel and other kernels, broadening its applicability. A notable recent application is in spatial transcriptomics, where the test is used to identify spatially varying genes. A practical challenge, however, is the selection of the kernel bandwidth (also called the range parameter in spatial statistics), which is often chosen empirically. Our recent theoretical results provide guidance on this choice. In this talk, I will give an overview of these results and their implications.

Key words: dependence, score tests, spatial models, spatial transcriptomics

Dynamic and Individualized Prediction of Cardiovascular Events: The International Childhood Cardiovascular Cohort (i3C) Consortium

Nanhua Zhang[1,2], Eric Odoom[3], Xia Wang[3], Jessica G. Woo[1,2]

[1] *Division of Biostatistics & Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA*

[2] *University of Cincicinnati College of Medicine, Cincinnati, OH, USA*

[3] *Division of Statistics & Data Science, Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH, 45040*

ABSTRACT

Our recent study has shown that cardiovascular risk factors of body-mass index, systolic blood pressure, total cholesterol level, triglyceride level and smoking beginning in early childhood, were associated with adult cardiovascular events and death from cardiovascular causes. In this article, we developed three joint models to predict the individual risk of CVD based on individual subject's demographic information and their CVD risk factors measured over time: 1) constant-coefficient joint model (CCJM), 2) varying coefficient joint model (VCJM), and P-spline joint model (mvJM-Spline). All three models have clinically useful results with AUC values of 0.84, 0.85, and 0.92, respectively. An R Shiny webtool was developed to implement the dynamic risk prediction tool.

Keywords: dynamic prediction, individualized prediction, cardiovascular disease

OPERA: An Interpretable Algorithm for Patient Stratification based on Partially Ordered Risk Factors

Menggang Yu[1,*]

Department of Biostatistics, University of Michigan

ABSTRACT

Risk stratification is an invaluable tool for modern healthcare systems. By separating patients into subgroups with distinct disease severity and prognosis, it allows better clinical decision making due to targeted care thus ultimately fosters healthier patient populations. In addition, it enhances communication, engagement, and research focus. This talk presents a new algorithm entitled 'Ordering Poset Elements by Recursive Amalgamation' (OPERA) for patient stratification using many risk factors. Health risk factors frequently exhibit total or partial ordering and when considered jointly, they form a partially ordered set or a poset. Compared with the usual regression models, OPERA can explore high order interactions similar to the well-known tree method. On the other hand, by explicitly exploring the poset structure, OPERA allows flexible and interpretable staging patterns and faster pruning for better performance. OPERA is evaluated in extensive simulation studies and cancer staging data to demonstrate its ability in performing risk stratification using ordered risk factors.

Keywords: Cancer staging, partial order, regression tree, risk stratification

α-Separability and Adjustable Combination of Amplitude and Phase Model for Functional Data

Tian Wang, **Jimin Ding**

Department of Statistics and Data Science, Washington University in Saint Louis

ABSTRACT

We consider separating and joint modelling amplitude and phase variations for functional data in an identifiable manner. To rigorously address this separability issue, we introduce the notion of α-separability upon constructing a family of α-indexed metrics. We bridge α-separability with the uniqueness of Fréchet mean, leading to the proposed adjustable combination of amplitude and phase model. The parameter α allows user-defined modelling emphasis between vertical and horizontal features and provides a novel viewpoint on the identifiability issue. We prove the consistency of sample Fréchet mean and variance, and proposed estimators. Our method is illustrated in simulations and COVID-19 infection rate data.

Keywords: Fréchet mean; functional data analysis; identifiability; joint model; separability; variation decomposition

Efficient Estimation for Recurrent Events under Informative Censoring Using Generalized Method of Moments

Yu-Jen Cheng[1] and Chang-Yu Tsai[2]

Institute of Statistics and Data Science, National Tsing Hua University

ABSTRACT

In this work, we develop semiparametric transformation models with a shared frailty variable for recurrent event data, accommodating correlation between the event process and censoring. Unlike standard shared-frailty proportional rate models, our framework allows nonproportional rate functions across covariates. Motivated by the decomposition of the rate function into shape and size components (Wang and Huang, 2014), we first adopt an inverse-rate weighting approach and then propose a generalized method of moments framework that integrates information from both components to improve efficiency. We establish their large-sample properties, evaluate finite- sample performance through simulations, and demonstrate practical utility with a real dataset.

Keywords: Generalized method of moments; Informative censoring; Recurrent event process

A Doubly Robust Instrumental Variable Approach for Estimating Average Treatment Effects in Time-to-Event Data with Unmeasured Confounding

Chung-Chou H. Chang[1,2,*], Runjia Li[2]

[1]*Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA*

[2]*Department of Biostatistics and Health Data Science, University of Pittsburgh School of Public Health, Pittsburgh, PA, USA*

ABSTRACT

We propose a novel doubly robust instrumental variable (IV) estimator for estimating average treatment effects (ATEs) in time-to-event outcomes subject to unmeasured confounding. While IV methods are increasingly applied in real-world data analyses, existing approaches for survival outcomes often rely on restrictive assumptions and lack desirable statistical properties. Our method is derived from the efficient influence function (EIF), ensuring double robustness and achieving asymptotic efficiency. The framework accommodates flexible machine learning estimators, making it well suited for complex electronic health record (EHR) data. Through extensive simulations, we demonstrate the method's robustness, asymptotic normality, and strong finite-sample performance. We apply the estimator to EHR data on ICU patients with vasopressor-dependent septic shock, using physician prescribing preference as the instrument to evaluate the effect of hydrocortisone on mortality. Results indicate no significant benefit or harm, providing reliable evidence for clinical decision-making. This doubly robust IV approach expands methodological tools for survival analysis under unmeasured confounding and enhances the validity of causal inference in real-world settings.

Keywords: Average treatment effect, efficient influence function, instrumental variable, time-to-event data, unmeasured confounding

A Quantile Cure Model with Partially Functional Covariate Effects

Chyong-Mei Chen[1], Yingwei Peng[2,*]

[1]*Institute of Public Health, School of Medicine, National Yang Ming Chiao Tung University*

[2]*Departments of Public Health Science and Mathematics and Statistics, Queen's University*

ABSTRACT

The quantile regression has several attractive features, such as its ability to allow covariate effects to vary at different quantile levels and to handle heteroscedasticity in data easily, which make it a viable alternative when analyzing data with continuous outcomes in recent years. In particular, it has been used in modeling survival data with and without a cured fraction. In this work, we propose novel estimating equation approaches to estimate a mixture cure model where the latency survival time is modeled by a quantile regression. The proposed estimation methods enjoy a double robustness in the sense that a misspecification in one of the two parts in the mixture cure model will not affect the estimation in the other part. The methods do not require the global log-linear assumption in the quantile regression, and they allow mixed effects of functional and constant effects in the regression when the log-linear assumption is hold in an interval of quantile levels. We established the asymptotic properties of the proposed estimators. Our simulation studies demonstrated the double robustness and the efficiency gains in the proposed estimators. An application of the proposed model and methods to data from a lung cancer study revealed new and interesting findings that were not reported in a previous analysis of the data.

Keywords: Estimating equation; Inverse probability censoring weight; Mixture cure model; Quantile regression model

Semiparametric Analysis of Multivariate Panel Count Data with Informative Observation Processes

Chang Chen[1], **Xin He[1,*]**

[1]*Department of Epidemiology and Biostatistics, University of Maryland, College Park*

ABSTRACT

Multivariate panel count data arise in studies involving several related types of recurrent events in which the study subjects are examined periodically over time. The observation times may vary from subject to subject and carry information about the underlying recurrent event processes of interest. In this paper, we propose a joint modeling approach to account for the informative observation processes using bivariate shared frailty models. Estimating equations and an EM algorithm are developed for the parameter estimation, and the resulting estimators are shown to be consistent and asymptotically normal. The proposed methods are evaluated through simulation studies and illustrated with an application to data from a skin cancer clinical trial.

Keywords: EM algorithm; Estimating equation; Informative observation times; Multivariate panel count data

Using Interpoint Distances to Develop a New Multivariate Control Chart Based on Change-Point Detection

Claudio Giovanni Borroni[1], Manuela Cazzaro[1], Paola Maddalena Chiodini[1]

[1] University of Milano - Bicocca (Italy)

ABSTRACT

The change-point paradigm has been successfully applied to statistical process control by building many parametric and nonparametric charts for sequential monitoring. We concentrate here on a novel implementation of the change-point paradigm, based on dynamic windows, which forces comparisons only between samples of the same size. Despite that implementation has been successfully applied to univariate control, some applications often need to consider more than a single variable at a time. Indeed, not only different aspects of quality must be considered, but their dependence structure can also provide relevant information. Thus, in this talk, we investigate the application of the dynamic-window approach to multivariate control variables. Specifically, we propose a preliminary reduction of the dimension of such variables, before the change-point methodology is applied. We evaluate some techniques of reduction based on inter-point distances. A comparison study is conducted to: i) identify techniques which are stable with respect to the actual dimension and shape of the underlying distribution; ii) identify techniques which can provide a fast signal when the underlying distribution undergoes shifts in location or scale. We focus on multivariate charts which perform averagely well across a wide number of cases, more than on those which are extremely powerful just in isolated cases.

Keywords: Multivariate statistical process control, Change-point detection, interpoint distances.

Some Change-Point Design-Based Distribution-free Approaches for Monitoring High-Dimensional Data

Amitava Mukherjee

Production, Operations and Decision Sciences Area, XLRI – Xavier School of Management, India.

ABSTRACT

This paper introduces some purely distribution-free schemes for monitoring high-dimensional data streams using a change point design setup, which combines the advantages of exponentially weighted moving averages. Proposed procedures are based on some distance measures and ranks. Our proposed procedures require collecting only a few reference samples from the in-control process at the outset of Phase II monitoring, and not many Phase I observations, unlike traditional Phase II distribution-free schemes for high-dimensional data. A regression-based approach is considered for determining the control limits. An industrial application in monitoring a process involving semiconductor quality is discussed. We also discuss comparative detection performance using Monte Carlo methods. The paper concludes with some remarks and recommendations for future research.

Keywords: Change Point Design; Distribution-free; High-dimensional; Process Monitoring

Large-Scale Decentralized Fault Diagnosis for Multi-Group Data with Auxiliary Information via Distributed Multiple Testing

Zhihan Zhang[1], Wendong Li[1], Fugee Tsung[2], Dongdong Xiang[1,*]

[1]*KLATASDS-MOE, School of Statistics, East China Normal University*

[2]*Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology*

ABSTRACT

In the era of big data, efficiently diagnosing faults in high-dimensional data streams (HDS) is critical for numerous industrial applications. This paper addresses a novel decentralized fault diagnosis problem involving multi-group HDS with multi-sequence auxiliary information (MAI). Traditional diagnostic methods, which are designed for a single group of single-sequence HDS, struggle with the complexity and volume of such data, often leading to suboptimal diagnostic performance. To overcome this challenge, we propose a distributed fault diagnosis framework that leverages advanced multiple testing techniques and data fusion strategies to analyze multi-group HDS with MAI. Under this framework, we introduce a generalized multi-sequence local index of significance for the data streams in each group, based on a Cartesian hidden Markov model, to effectively fuse information from auxiliary sequences. This is then integrated into a distributed multiple testing procedure for group-wise diagnosis of the target data sequence. The proposed procedure minimizes the group-wise expected number of false positives in the target sequence while controlling the overall group-wise missed discovery rate at a specified level. Numerical studies demonstrate that the proposed method outperforms state-of-the-art diagnostic techniques, providing more reliable and effective fault diagnostics.

Keywords: Auxiliary Information; Distributed Computing; Fault Diagnosis; Multi-Group Data; Multiple Testing

Control Chart for High-Dimensional Dynamic Process Monitoring

Peihua Qiu

University of Florida

ABSTRACT

In air pollution surveillance, additive manufacturing, and other fields, monitoring high-dimensional data streams presents unique challenges. The in-control (IC) distributions often change over time due to seasonality and other factors, making it difficult to apply standard statistical process control (SPC) charts. Traditional SPC methods assume that IC process observations at different times are independent and identically distributed—an assumption that is often invalid in these settings. In this talk, we present a novel process monitoring method that integrates principal component analysis with sequential learning. This approach effectively handles high dimensionality, time-varying IC distributions, serial data correlation, and nonparametric data distributions. It has proven to be a reliable analytic tool for the online monitoring of high-dimensional dynamic processes. This is joint research with Dr. Xiulin Xie.

Keywords: Control charts; Correlation; Dynamic processes; Online monitoring

Survival Models with a Cured Fraction: A Zero-Inflated Gamma Frailty–Copula Approach

Masaki Hino[1,2], Takeshi Emura[3,4], Shogo Kato[1]

[1]*The Institute of Statistical Mathematics*

[2]*The Graduate University for Advanced Studies, SOKENDAI*

[3]*School of Informatics and Data Science, Hiroshima University*

[4]*Biostatistics Center, Kurume University*

ABSTRACT

We introduce a zero-inflated gamma frailty copula model to analyze dependent survival times in the presence of long-term survivors. Existing cure frailty copula approaches typically model frailty as a discrete random variable with a mass at zero, so that survivors are naturally represented by zero frailty. While this captures the cured fraction, it offers limited flexibility for modeling continuous heterogeneity among susceptible individuals. By combining a point mass at zero with a continuous gamma component, our construction simultaneously represents long-term survivors and heterogeneous frailty within a copula framework.

The proposed bivariate survival function can be represented as a mixture distribution that depends on whether each subject is cured or not. To estimate the parameters under this mixture structure, we employed both an expectation–maximization (EM) algorithm and a Newton-type algorithm within the zero-inflated gamma frailty copula framework. Numerical experiments indicate that these approaches yield maximum likelihood estimators that provide reasonable estimates of the model parameters. Finally, application to a real clinical dataset demonstrates how the proposed model can effectively capture both the cured fraction and the dependence structure of paired survival times.

Keywords: cure model; zero-inflated-gamma distribution; frailty-copula model; EM algorithm

Mean Residual Life Based Illness-Death Model for Semicompeting Risks Data

Huang Rui, **Liming Xiang**

School of Physical and Mathematical Sciences, Nanyang Technological University

ABSTRACT

Semicompeting risks data are available in many studies, where a nonterminal event (e.g., disease progression) is of interest and may be censored by the occurrence of a terminal event (e.g., death). The illness-death model has been developed as a common approach for regression analysis of such data by modelling transitions between three states using the proportional hazards or accelerated failure time model. In this work, we propose an illness-death model formulated using the mean residual life function, offering a straightforward and interpretable modelling framework for state transitions in the semicompeting risks setting. We facilitate estimation through novel estimating equations derived from a penalized quasi-likelihood approach incorporating inverse probability weighting. Unlike the conventional illness-death model assuming a shared gamma frailty, our method requires no distributional assumption on the latent frailty term, thereby reducing the risk of model misspecification. Simulation studies demonstrate its promising performance in across a range of realistic scenarios and an application to a myeloma progression study illustrates its practical utility.

Keywords: Buckley-James estimate; Conditional inference; Dependent censoring; Inverse probability of censoring weight; Quasi-likelihood function

H-Likelihood Approach on the Joint Frailty Model for Clustered Bivariate Survival Data

Jihoon Kwon[1], Jia-Han Shih[2], Takeshi Emura[3], **Il Do Ha[1]**

[1]*Department of Statistics & Data Science, Pukyong National University, Busan, South Korea*

[2]*Department of Applied Mathematics, National Sun Yat-sen University, Kaohsiung, Taiwan*

[3]*School of Informatics and Data Science, Hiroshima University, Hiroshima, Japan*

ABSTRACT

Recently, clustered survival data have been extensively studied using various correlated modelling approaches, such as frailty models and copula models. These data can take several forms, including bivariate censored data, semi-competing risks data, and competing risks data. Traditionally, each type has been analyzed using a separate model. In this talk, we propose a unified joint frailty modelling approach which is capable of handling all three types of clustered survival data within a single model-based likelihood framework. Here, the unknown baseline hazards in the joint frailty models are modeled based on a cubic M-spline basis function that does not require a specific parametric form. Inference for the model parameters is performed via the hierarchical likelihood (h-likelihood; Lee and Nelder, 1996) method, which avoids the intractable integration over frailty required in marginal likelihood approaches and effectively captures heterogeneity across clusters. Unlike the classical likelihood for fixed parameters only, the h-likelihood is constructed for both fixed parameters and unobserved frailties at the same time. The performance of the proposed approach is evaluated through simulation studies, which demonstrate that the estimated regression coefficients appear reasonable for all three types of survival data. The proposed method is further illustrated using three real-world datasets.

Keywords: Clustered bivariate survival data; Competing-risks data; Joint frailty models; H-likelihood

Inferring Median Survival under Dependent Censoring

Takeshi Emura

School of Informatics and Data Science, Hiroshima University, Japan

ABSTRACT

The key difficulty in survival analysis is the proper handling of censoring. So far, existing inference methods for median survival have been developed under the independent censoring assumption, which is too strong for many applications. As a solution, we develop new methods for dependent censoring regimes. This is accomplished by the median estimator from copula-graphic estimators developed for survival copula models. The proposed method is a median version of our previously proposed method for the Mann-Whitney effect [1]. We present this methodology with simulations and data examples. This is the joint work with Dennis Dobler.

Reference:
[1] Emura, T., Ditzhaus, M., Dobler, D., & Murotani, K. (2024). Factorial survival analysis for treatment effects under dependent censoring. Statistical Methods in Medical Research, 33(1), 61-79.

Keywords: Archimedean copula; Copula; Copula-graphic estimator; Survival analysis; Treatment effect

Degradation Models for Life Time Estimation of Serial and Parallel Connected Lithium-ion Battery Packs

Shuen-Lin Jeng, Yi-No Tseng

Department of Statistics and Institute of Data Science, National Cheng Kung University

ABSTRACT

This study utilized capacity loss datasets from Lithium-ion battery cells and packs to calculate the State of Health (SoH) after each discharge cycle. To more accurately depict the degradation paths of cell batteries within a pack and of the pack itself, we introduced the concept of the discharge rate factor to describe the impact of variations in discharge rates during each discharge cycle. This factor was incorporated into a random coefficient degradation model. We proposed several Cell-to-Pack methods to estimate the reliability of serial and parallel connected battery packs by using the cell data. The results indicate that compared to traditional reliability methods, such as the Reliability Block Diagram, our Cell-to-Pack methods yield a more accurate battery pack reliability estimation

Keywords: Cell-to-Pack methods; random coefficient degradation model; reliability; State of Health

Optimal Designs for Gamma Degradation Tests

Hung-Ping Tung[1], Yu-Wen Chen[1]

Department of Industrial Engineering and Management, National Yang Ming Chiao Tung University

ABSTRACT

This study analytically investigates the optimal design of gamma degradation tests, including the number of test units, the number of inspections, and inspection times. We first derive optimal designs with periodic inspection times under various scenarios. Unlike previous studies that typically rely on numerical methods or fix certain design parameters, our approach provides an analytical framework to determine optimal designs. In addition, the results are directly applicable to destructive degradation tests when number of inspection is one. The investigation is then extended to designs with aperiodic inspection times, a topic that has not been thoroughly explored in the existing literature. Interestingly, we show that designs with periodic inspection times are the least efficient. We then derive the optimal aperiodic inspection times and the corresponding optimal designs under two cost constraints.

Keywords: Reliability; Degradation tests; Gamma process; Inspection time; Optimal design

Shrinkage Estimation for the Rate Parameter under the Exponential Distribution with Censored Survival Data

Nanami Taketomi[1], Kosuke Nakazono[2], Akane Okada[3], and Takeshi Emura[4]

[1] *School of Information and Data Sciences, Nagasaki University*

[2] *Department of Industrial Engineering and Economics, School of Engineering, Institute of Science Tokyo*

[3] *Department of Biostatistics, Graduate school of medicine, Kurume University*

[4] *School of Informatics and Data Science, Hiroshima University*

ABSTRACT

The exponential distribution is widely applied in survival and reliability analyses. To estimate the rate parameter of the exponential distribution with censored survival data, maximum likelihood estimation is usually employed. However, in case the number of events or subjects is small for the data, the maximum likelihood estimator (MLE) can have large bias and variance. In this talk, we present a shrinkage estimator for the rate parameter of the exponential distribution using a penalized log-likelihood by adding a penalty for a large parameter value. The value of the shrinkage parameter is selected by maximizing the likelihood cross validation. We also derive theoretical properties of the penalized maximum likelihood estimator. Simulation results show that the proposed method provides the smaller mean squared error than the maximum likelihood estimation. Finally, we apply the proposed method to the reliability data and the prostate cancer data. In the reliability dataset, failure events account for only 0.25% of the total observations, indicating a highly censored data structure. The prostate cancer data contains 7.2% events of the total observations. In the analysis, the reliability functions by the MLE and the shrinkage estimator were drawn. The reliability function by the PML estimator is expected to have the smaller MSE than the ML estimator.

Keywords: Exponential distribution, Maximum likelihood estimation, Mean squared error, Shrinkage estimation, Survival data

Reliability Analysis for Small Ball Bearings by Considering the Correlation of Their Lifetimes

Shuhei Ota[1]

[1]*Department of Industrial Engineering and Management, Kanagawa University, Japan*

ABSTRACT

Ball bearings are crucial components in rotating machinery, and their failure can have a significant impact on the reliability of the machinery. Although many life test rigs for bearings have been developed to assess their reliability, few studies have focused on small bearings or multiple bearings tested simultaneously, where dependent failures may occur due to shared loads. This study extends our previous work on developing a life test rig for two small ball bearings (inner diameter 10 mm) operated under radial load and monitored by acceleration and temperature sensors. We conducted accelerated life tests to investigate the correlation between the bearings' lifetimes. Statistical analysis of temperature data revealed both positive and negative correlations between the two bearings, suggesting the presence of dependent failures due to load transfer and shaft wear. Lifetime data were fitted using Weibull, Gamma, and lognormal distributions, with the lognormal model showing the best fit. Based on the estimated dependence structure, a reliability model considering correlated lifetimes was formulated. The proposed analysis provides insights into how the interaction between components affects overall system reliability. The results contribute to the development of more accurate methods for predicting the remaining useful life of systems with multiple dependent components.

Keywords: Reliability analysis; Condition monitoring data; Dependence; Copula

A Simple Nonparametric Least-Squares-Based Causal Inference for Heterogeneous Treatment Effects

Ying Zhang[1*], Yuanfang Xu[2], Lili Tong[1], Giorgos Bakoyannis[3], and Bin Huang[4]

[1] *Department of Biostatistics, University of Nebraska Medical Center*

[2] *Bristol Myers Squib*

[3] *Department of Biostatistics and Health Data Science, Indiana University*

[4] *Cincinnati Children's Hospital Medical Center*

ABSTRACT

Estimating treatment effects is a common practice in making causal inferences. However, it is a challenging task for observational studies because the underlying models for outcome and treatment assignment are unknown. The concept of potential outcomes has been widely adopted in the literature on causal inference. Building on potential outcomes, we propose a simple nonparametric least-squares spline-based causal inference method to estimate heterogeneous treatment effects in this manuscript. We use empirical process theory to study its asymptotic properties and conduct simulation studies to evaluate its operational characteristics. Based on the estimated heterogeneous treatment effects, we further estimate the average treatment effect and show the asymptotic normality of the estimator. Finally, we apply the proposed method to assess the biological anti-rheumatic treatment effect on children with newly onset juvenile idiopathic arthritis disease using electronic health records from a longitudinal study at Cincinnati Children's Hospital Medical Center.

Keywords: Causal inferences; Empirical process theory; Heterogeneous treatment effects; Potential outcome; Regression splines

Using Negative Controls to Adjust for Unmeasured Confounding in Continuous Exposure Settings

Kate Hu[1], Dafne Zorzetto, Francesca Dominici

Department of Statistics, The Ohio State University

ABSTRACT

This talk introduces approaches for using negative controls to adjust for unmeasured confounding in observational dose–response studies where both exposure and outcome are continuous. I will present several machine learning techniques and a Bayesian nonparametric method we developed to estimate the causal exposure–response function (CERF) using negative control information. Our Bayesian nonparametric method models the CERF as a mixture of linear models, enabling flexibility to capture nonlinear patterns while preserving computational efficiency and benefiting from closed-form results under linear assumptions. I will share simulation studies evaluating these methods' performance. As an illustration, I will show how to select negative controls and use our open-source tools to assess the relationship between long-term ambient $PM_{2.5}$ exposure and cardiovascular hospitalization rates among older adults in the continental United States, accounting for potential unmeasured confounders. Finally, I will discuss ongoing methodological improvements and alternative estimands to consider for causal inference with continuous exposures.

Keywords: bias; confounding; negative controls; Bayesian analysis

Causal Mediation Analysis for Survival Outcome and Recurrent Event Mediators with Time-Varying Confounding

Fang Niu[1], **Cheng Zheng[1,*]**, Lei Liu[2]

[1]*Department of Biostatistics, University of Nebraska Medical Cener*

[2]*Division of Biostatistics, Washington University in St. Louis*

ABSTRACT

Recurrent events and repeated measures are commonly encountered in clinical longitudinal studies, often holding strong associations with patient outcomes. Although joint models for repeated measure, recurrent events, and a terminal event have been developed to account for their correlation, limited methodologies exist to rigorously examine causal mediation mechanisms involving multiple types of intermediate time-varying variables, especially when these variables are causally related. This study addresses this gap by proposing a novel causal mediation analysis framework to quantify natural direct and indirect effects when the recurrent events and survival outcome are confounded by longitudinal measured time-varying variable. We extend joint modeling approaches by incorporating shared random effects (frailties) structures, relaxing the commonly used "sequential ignorability" assumption, and accounting for unmeasured time-independent confounders through shared random effects. Simulation studies demonstrate the robustness and finite sample performance of our estimators for natural direct and indirect effects. We apply our method to the Terry Beirn Community Programs for Clinical Research on AIDS (CPCRA) study and demonstrate that recurrent opportunistic infections (OIs) mediate the effects of prior AIDS-defining conditions on survival outcomes after taking into account the potential confounding effect of time-varying CD4 count measurements.

Keywords: Causal mediation analysis; Joint modeling; Recurrent event; Repeated measurement

Bayesian Causal Discovery with Cycles and Latent Confounders

Wei Jin[1], Lang Lang[1], Leah H. Rubin[2], **Yanxun Xu[1]**

[1]*Department of Applied Mathematics and Statistics, Johns Hopkins University*

[2]*Department of Neurology, School of Medicine, Johns Hopkins University*

ABSTRACT

Learning causality from observational data has received increasing interest across various scientific fields. However, most existing methods assume the absence of latent confounders and restrict the underlying causal graph to be acyclic, assumptions that are often violated in many real-world applications. In this paper, we address these challenges by proposing a novel framework for causal discovery that accommodates both cycles and latent confounders. By leveraging the identifiability results from noisy independent component analysis and recent advances in factor analysis, we establish the unique causal identifiability of the proposed method under mild conditions. We further develop a fully Bayesian approach for causal structure learning and evaluate its identifiability, utility, and superior performance against state-of-the-art alternatives through extensive simulation studies. Application to a dataset from the Women's Interagency HIV Study yields interpretable and clinically meaningful insights. To facilitate broader applications, we have implemented the proposed Bayesian causal discovery method in an R package, \pkg{BayCausal}, which is the first publicly available software capable of achieving unique causal identification in the presence of both cycles and latent confounders.

Keywords: Bayesian structural learning; Causal identification; Directed cyclic graph; Latent confounding; Observational data

Bayesian Automated Learning of Sparsity in Risk Prediction with Application to Whole-brain Functional Connectivity Analysis

Taiwo Fagbohungbe, Eric Odoom, **Xia Wang**, Xuan Cao

Division of Statistics and Data Science, Department of Mathematical Sciences, University of Cincinnati, USA

ABSTRACT

Challenges in disease risk models lie in identifying key biomarkers and estimating their associated coefficients. Many existing approaches rely on prespecified tuning parameters that implicitly control the number of relevant variables, the so-called sparsity level. We propose a fully Bayesian framework based on spike-and-slab priors for logistic risk prediction that automatically learns sparsity from the data. Specifically, we place hierarchical priors on the variable inclusion probability and on the slab variance (shrinkage) parameter. Learning both parameters enables automated adaptation to the true model sparsity, yielding flexible yet accurate coefficient estimates and predictions. Under mild conditions, the method attains model selection consistency with computational complexity comparable to existing approaches. We analyze an Autism dataset for risk stratification using whole-brain functional connectivity features, training on the NYU site and evaluating on an independent UCLA site. The proposed model achieves 81% classification accuracy, outperforming competing methods. We further evaluate its performance on a mouse gene expression dataset, attaining 75% accuracy and surpassing strong baselines.

Keywords: fMRI, Hierarchical model, Logistic regression, Risk prediction, Spike-and-slab

Bayesian Design and Analysis Methods for Decentralized Clinical Trials

Ruitao Lin

The University of Texas MD Anderson Cancer Center

ABSTRACT

Decentralized clinical trials (DCTs) extend trial activities beyond traditional sites, improving access, convenience, efficiency, and result generalizability. They are particularly promising for chronic conditions like diabetes and obesity, which require longer study durations to assess drug effects. However, decentralized data collection raises concerns about increased variability and potential biases. In this talk, I will present several novel Bayesian methodologies developed at MD Anderson Cancer Center in collaboration with researchers from major pharmaceutical companies to address the design and analysis of decentralized clinical trials with longitudinal data. In particular, I will discuss a novel Bayesian integrated learning procedure that combines centralized and decentralized data collection for analyzing longitudinal data in DCTs. Through simulations and sensitivity analyses, we demonstrate that the proposed Bayesian integrated learning method performs well across various scenarios. Notably, it matches the efficiency of traditional trials when decentralized data collection introduces no additional variability or error. Even when such issues arise, it remains less biased and more efficient than naïve methods that rely solely on centralized data or indiscriminately pool data from both sources.

Keywords: Bayesian methods, clinical trials, decentralization, measurement errors.

Bayesian Dose-Response Meta-Analysis for Predictive Biomarkers Using Aggregate and Individual Participant Data with Data Augmentation for Precision Medicine

J. Jack Lee and Wayne Yi-Hung Wu

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, USA

ABSTRACT

A predictive biomarker is a biological indicator that identifies individuals who are more likely than those without the biomarker to experience either a positive or negative effect from a medical product. It plays a pivotal role in precision medicine. For example, high PD-L1 expression is associated with high efficacy for immune checkpoint inhibitors. In reported studies, some may have individual participant data (IPD) and many others have aggregate data (AD) but may have varying biomarker cutoff values from studies to studies.

To synthesize predictive biomarker evidence from studies with different data types and cutoff values, we construct a Bayesian hierarchical dose–response meta-analysis for time-to-event outcomes that integrates both IPD and AD through a four-parameter log-logistic model. For AD-only settings, a data-augmentation method using quasi–Monte Carlo integration is proposed to average over the biomarker distribution within each interval. Simulation studies demonstrate that incorporating IPD and prognostic factor adjustment substantially reduced bias and mean-squared error, while the augmented AD model improved slope estimation, especially under steep nonlinear relationships. In addition, we introduce a covariate-adjusted meta-analytic combined (cMAC) framework for augmenting the control arm of a current randomized trial by borrowing from historical controls available as IPD, AD, or both. The model links aggregate-level likelihoods to individual-level models via quasi–Monte Carlo integration, enabling coherent borrowing while accounting for covariate imbalance. Simulation studies across varying heterogeneity and covariate-shift scenarios show that hybrid IPD + AD borrowing improves precision and power relative to no borrowing, while automatically downweighting incompatible data to control bias.

Keywords: Bayesian hierarchical model, log-logistic dose–response model, meta-analysis, data augmentation, covariate-adjusted meta-analytic combined framework

December 18 (Thursday):

Parallel Sessions [12:50 – 14:30]:

18a1 - Recent Developments for Data Science

18a2 - Statistical Innovations for High-dimensional and Time-Dependent Data

18a3 - Analysis of Complex Data

18a4 -

18a5 - New Data Analytics for Evaluating Complex Associations

18a6 - Causal Inference for Survival Data

18a7 - Recent Advances in Time Series Analysis

18a8 - 3D Protein Structure InformaticsAnalysis

18a9 - New Developments in High-Dimensional Matrix and Network Analysis

18a10 - Recent Methods Development in High Dimensional Omics Studies

Parallel Sessions [14:50 – 16:30]:

18b1 - Spatial and Environmental Statistics

18b2 - New Insights into Inference and Causality

18b3 - Recent Developments for Biomedical Statistics

18b4 - Recent Advances in Optimal Experimental Designs

18b5 - High Dimensional Models with Applications in Biomedical Sciences

18b6 - Modern Machine Learning in the Big Data Era

18b7 - Classical Meets Cutting-Edge: Regression, Mixtures, and Joint Models in Biomedical Research

18b8 - Exploring Phenomena Through Mathematical Modeling

18b9 - Copula and Dependence Modeling

18b10 - Modern Bayesian Tools for Modeling and Inference

AI, BI & SI—Artificial, Biological and Statistical Intelligences

Dennis K.J. Lin

Distinguished Professor, Department of Statistics, Purdue University, West Lafayette, IN

ABSTRACT

Artificial Intelligence (AI) is clearly one of the hottest subjects these days. Basically, AI employs a huge number of inputs (training data), super-efficient computer power/memory, and smart algorithms to perform its intelligence. In contrast, Biological Intelligence (BI) is a natural intelligence that requires very little or even no input. This talk will first discuss the fundamental issue of input (training data) for AI. After all, not-so-informative inputs (even if they are huge) will result in a not-so-intelligent AI. Specifically, three issues will be discussed: (1) input bias, (2) data right vs. right data, and (3) sample vs. population. Finally, the importance of Statistical Intelligence (SI) will be introduced. SI is somehow in between AI and BI. It employs important sample data, solid theoretically proven statistical inference/models, and natural intelligence. In my view, AI will become more and more powerful in many senses, but it will never replace BI. After all, it is said that "The truth is stranger than fiction, because fiction must make sense." The ultimate goal of this study is to find out "how can humans use AI, BI, and SI together to do things better."

Keywords: Data Quality, Data Right and Right Data, Design of Experiment, Intelligent Data Collection Method

Solving the Mysteries of Place Cells and Grid Cells by Representation Learning

Ying Nian Wu
UCLA

ABSTRACT

The 2014 Nobel Prize in Physiology or Medicine recognized the discovery of place cells and grid cells in the mammalian brain. Each place cell fires at a single specific location, whereas each grid cell fires at multiple locations forming a hexagonal grid pattern. Yet the computational principles underlying these phenomena have remained mysterious. We show both emerge from representation learning through geometric optimization. Grid cells learn embeddings that preserve local distances through conformal isometry, forming a coordinate system. We prove hexagonal patterns are optimal: hexagonal flat tori uniquely minimize deviation from local distance preservation by distributing curvature isotropically through six-fold symmetry. Building upon this coordinate system, place cells learn embeddings that preserve spatial adjacency relations defined by transition kernels of heat diffusion with reflecting boundary conditions, thereby forming a cognitive map. Specifically, inner products between embeddings reconstruct transition probabilities, causing localized firing patterns to emerge automatically from non-negative matrix factorization constraints. This reveals how the brain solves navigation by transforming spatial reasoning into optimization on learned geometric representations.

Keywords: representation learning, computational neuroscience, spectral decomposition, non-negative matrix factorization, sparse coding.

Generate Diverse Protein Conformations through AlphaFold

Samuel Kou[1]

Department of Statistics, Harvard University

ABSTRACT

The introduction of AlphaFold has revolutionized the task of protein structure prediction from a given sequence of amino acids; the groundbreaking contribution of AlphaFold was recognized by the 2024 Nobel Prize in Chemistry. As a deep-learning based method, AlphaFold was trained from the publicly available Protein Data Bank (PDB), a database of known protein structures. An inherent limitation of AlphaFold is that its prediction can only give a static structure, whereas in reality, the structures of proteins are dynamic and can change in response to their environment or binding partners, with significant biological consequences. In this talk, we focus on enhancing and diversifying protein structure prediction using AlphaFold. Through a principled iterative statistical sampling framework, we significantly expand AlphaFold's capabilities, enabling it to explore a broader conformational space. Key methodologies involve modifying the multiple sequence alignment (MSA) and template inputs to encourage AlphaFold to explore different conformations, thereby increasing structural diversity. This is achieved in particular through an iterative sequential sampling approach, which allows for the incorporation of protein residue co-evolutionary information in the structure prediction, broadening the conformational possibilities that AlphaFold can investigate. We will illustrate the capabilities of the statistical sampling approach through examples.

Keywords: Protein folding; sequential sampling; coevolutionary information; protein conformation

Recent Advances in MM Optimization Algorithms

Hua Zhou[1], Xunjian Li[1], Kenneth Lange[2,3]

[1]*Department of Biostatistics, University of California, Los Angeles*

[2]*Department of Human Genetics, University of California, Los Angeles*

[3]*Department of Computational Medicine, University of California, Los Angeles*

ABSTRACT

The majorization-minimization (MM) principle is an extremely general framework for deriving optimization algorithms. It includes the expectation-maximization (EM) algorithm, proximal gradient algorithm, concave-convex procedure, quadratic lower bound algorithm, and proximal distance algorithm as special cases. Besides numerous applications in statistics, optimization, and imaging, the MM principle finds wide applications in large-scale machine learning problems such as matrix completion, discriminant analysis, and nonnegative matrix factorizations. This talk presents some novel applications of the MM principle in the big data setting, including parallel block least squares, deweighting weighted least squares, large-scale variance component model, independent component analysis, and multi-level Monte Carlo.

Keywords: majorization-minimization; optimization.

Asymptotic FDR Control with Model-X Knockoffs: Is Moments Matching Sufficient?

Yingying Fan[1], Lan Gao, Jinchi Lv, Xiaocong Xu[2,*]

Data Sciences and Operations, University of Southern California

ABSTRACT

We propose a unified theoretical framework for studying the robustness of the model-X knockoffs framework by investigating the asymptotic false discovery rate (FDR) control of the practically implemented approximate knockoffs procedure. This procedure deviates from the model-X knockoffs framework by substituting the true covariate distribution with a user-specified distribution that can be learned using in-sample observations. By replacing the distributional exchangeability condition of the model-X knockoff variables with three conditions on the approximate knockoff statistics, we establish that the approximate knockoffs procedure achieves the asymptotic FDR control. Using our unified framework, we further prove that an arguably most popularly used knockoff variable generation method--the Gaussian knockoffs generator based on the first two moments matching--achieves the asymptotic FDR control when the two-moment-based knockoff statistics are employed in the knockoffs inference procedure. For the first time in the literature, our theoretical results justify formally the effectiveness and robustness of the Gaussian knockoffs generator. Simulation and real data examples are conducted to validate the theoretical findings.

Keywords: Model-X knockoffs; Gaussian knockoffs generator; moments matching; asymptotic FDR control; robustness

Detection of Dynamic Instability by Dispersion Ratios in Local Block Lyapunov Exponent Diagrams

Rintaro Ichiya[1], Rinka Sagawa[1], **Yan Liu[2,*]**

[1]*Department of Applied Mathematics, Waseda University*

[2]*Faculty of Science and Engineering, Waseda University*

ABSTRACT

This paper presents statistical methods for quantifying local dynamic instability arising from chaotic behaviors in stochastic processes. We introduce the local block Lyapunov exponent and the diagonal Lyapunov dispersion ratio as fundamental statistical tools to distinguish chaotic behaviors in stochastic processes. The diagonal Lyapunov dispersion ratio is used as a macroscopic measure to investigate the distributional distortion in each block of stochastic processes. We develop the asymptotic theory for these statistical tools under a general setting. Numerical simulations under different parameter settings illustrate the satisfactory performance of our statistical approach. We also apply this method to the financial market data, providing evidence for the possible local dynamic instability in the data.

Keywords: Dynamic instability; Diagonal Lyapunov dispersion ratio; Local block Lyapunov exponent; Nonlinear time series

Variable Selection for High-Dimensional Heteroscedastic Regression and Its Applications

Po-Hsiang Peng[1], Hai-Tang Chiou[2], **Hsueh-Han Huang[3]**, Ching-Kang Ing[1]

[1]*Institute of Statistics and Data Science, National Tsing Hua University, Hsinchu, Taiwan*

[2]*Department of Mathematics, National Chung Cheng University,Chiayi, Taiwan*

[3]*Institute of Statistical Science, Academia Sinica, Taipei, Taiwan*

Abstract

We are examining variable selection in high-dimensional linear heteroscedastic models. Drawing inspiration from the connection between the linear heteroscedastic function and the interaction model, we develop a two-stage algorithm to identify the relevant variables in the model mentioned above. We demonstrate the selection consistency of our proposed two-stage method and highlight its efficacy through numerical simulations. Furthermore, we leverage our method to pinpoint defective tools during the semiconductor manufacturing process.

Keywords: High-dimensional; interaction model; Linear heteroscedasticity; Model selection; Multiplicative heteroscedasticity

Adaptive High-Dimensional Model Selection via Chebyshev's Greedy Algorithm

Chien-Tong Lin[1], **Chi-Shian Dai[2]**, You-Lin Chen[3], Ching-Kang Ing[4]

Department of Applied Mathematics, National Sun Yat-sen University

Department of Statistics and Data Science, National Cheng Kung University Amazon.com, Inc.

Insititute of Statistics and Data Science, National Tsing Hua University

ABSTRACT

Sparsity assumptions on regression coefficients play a central role in high-dimensional model selection. However, the sparsity level is typically unknown in practice, motivating the development of procedures that perform robustly across a range of sparsity regimes. In this paper, we investigate the convergence behavior of Chebyshev's Greedy Algorithm (CGA) under varying sparsity levels and propose selecting the number of CGA iterations using a high-dimensional information criterion (HDIC). We show that the resulting procedure, CGA combined with HDIC (CGA+HDIC), is adaptive in the sense that it automatically achieves the optimal trade-off between variance and squared bias without prior knowledge of the sparsity level. As a key application, we demonstrate that CGA+HDIC attains the optimal convergence rate (up to a logarithmic factor in the sample size) in high-dimensional generalized linear models. Theoretical results are supported by extensive simulation studies and real data analyses.

Keywords: Chebyshev greedy algorithm; generalized linear models; high-dimensional information criterion; sparsity levels

Statistical Thinking and AI Transformers: A Two-Way Exchange Between Time Series and Attention Mechanisms

Jiecheng Lu (student), **Shihao Yang**

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

ABSTRACT

Despite Transformers' success across AI domains, their application to time series forecasting remains lukewarm. For example, in infectious disease forecasting it showed no clear advantage over classical statistical baselines, motivating us to bridge attention mechanisms and classical time series principles.

We first develop a statistical account of when and why attention should work: a single linear attention layer behaves like a low-rank Vector Autoregression (VAR), while stacking layers induces higher-rank lag interactions, yielding an attention design aligned with VAR structure. We then observe that typical Transformers are autoregressive only, missing the moving-average component in statistical time series models. By introducing an attention pathway over residuals inspired by ARMA models, we improve forecasting accuracy. For efficiency and long contexts, we reinterpret linear attention as a truncated softmax and add dedicated pathways (beyond Q, K, V) capturing higher-order Taylor series to recover expressivity at linear time. We show how "prompt engineering" for time series enables zero-shot and few-shot forecasting through in-context learning, how auxiliary "scratch-paper" channels act as a chain-of-thought analogue for multivariate series, and how targeted training techniques stabilize and accelerate convergence.

Together, these contributions demonstrate how statistical thinking enables better Transformer designs for time series and reciprocally, how time series insights yield improved attention mechanisms that transfer to vision and text domains. This two-way exchange between classical statistical principles and modern deep learning architectures opens new directions for both fields.

Keywords: Time Series; Attention Mechanisms; Transformers; AI Foundation Models

HeteroJIVE: Joint Subspace Estimation for Heterogeneous Multi-View Data

Jingyang Li[1], Zhongyuan Lyu[2]

[1]*University of Michigan, Ann Arbor*

[2]*The University of Sydney*

ABSTRACT

Many modern datasets consist of multiple related matrices measured on a common set of units, where the goal is to recover the shared low-dimensional subspace. While the Angle-based Joint and Individual Variation Explained (AJIVE) framework provides a solution, it relies on equal-weight aggregation, which can be strictly suboptimal when views exhibit significant statistical heterogeneity (arising from varying SNR and dimensions) and structural heterogeneity (arising from individual components). In this paper, we propose HeteroJIVE, a weighted two-stage spectral algorithm tailored to such heterogeneity. Theoretically, we first revisit the ``non-diminishing'' error barrier with respect to the number of views K identified in recent literature for the equal-weight case. We demonstrate that this barrier is not universal: under generic geometric conditions, the bias term vanishes and our estimator achieves the $O(K^{-1/2})$ rate without the need for iterative refinement. Extending this to the general-weight case, we establish error bounds that explicitly disentangle the two layers of heterogeneity. Based on this, we derive an oracle-optimal weighting scheme implemented via a data-driven procedure. Extensive simulations corroborate our theoretical findings, and an application to TCGA-BRCA multi-omics data validates the superiority of HeteroJIVE in practice.

Keywords: JIVE, heterogeneity, Multi-view data analysis

A Riemannian Factor Model for Manifold-valued Time Series

Shuo-Chieh Huang[1], Rong Chen[1], Yaqing Chen[1]

[1]*Department of Statistics, Rutgers University*

ABSTRACT

In this paper, we propose the Riemannian factor model, a novel framework for analyzing time series taking values in Riemannian manifolds in potentially high dimensions. Such time series is encountered in many applications, including economics, finance, medical imaging, and genomics and microbiome research. The proposed model is geometry-aware and accounts for the inherent nonlinearity in the data. Under a high-dimensional asymptotic regime, where the manifold dimension is allowed to diverge with n, the sample size, we establish convergence rates for the estimated loading space. In particular, under short-memory and strong factor conditions, we obtain a dimension-free n^{-1/2} rate, which matches the convergence rates of the fixed-dimensional Riemannian principal component analysis and the high-dimensional linear factor models with strong factors. Applied to the covariance matrices of selected U.S. stock returns, viewed as time series in the Bures-Wasserstein manifold, the proposed method yields interpretable factors and competitive predictions.

Keywords: dimension reduction, non-Euclidean time series; compositional time series; Bures-Wasserstein metric; covariance prediction.

Enhancing Generalizability and Fairness of HIV Risk Predictions: A Machine Learning Approach Using EHR Data

Hulin Wu

The Betty Wheless Trotter Professor, Department of Biostatistics & Data Science,

School of Public Health, University of Texas Health Science Center at Houston

ABSTRACT

Despite significant progress in HIV prevention and treatment, underdiagnosis remains a major driver of new infections. We developed and validated a machine learning–based HIV-1 risk prediction model using the Cerner Health Facts® nationwide electronic health record (EHR) database, encompassing more than 69 million patients across 85 U.S. health systems (2000–2018). An automated HIV phenotyping algorithm identified 38,310 HIV-1 cases and 118,090 controls, from which 4,442 candidate predictors were screened. A refined LASSO logistic regression model retained 281 predictors and achieved high discrimination (AUC = 0.927; sensitivity = 79.2%; specificity = 92.7%). To address fairness, we evaluated performance across demographic subgroups and implemented a threshold-based post-processing method to mitigate disparities in false-positive and true-positive rates, particularly across race groups. This study is the first to systematically integrate algorithmic fairness into HIV-1 prediction, demonstrating that EHR-based models can effectively and equitably identify underdiagnosed individuals. Our approach offers a scalable framework for improving HIV case finding and supporting equitable clinical decision-making in diverse healthcare settings. This is the work with my students, Tianheng Zhang and Yuxuan Gu.

Keywords: AIDS; EMR; Machine Learning Fairness

Examining Directional Association between Depression and Anxiety in US Medical Interns

Soumik Purkayastha[1], Peter X.-K. Song[2,*]

[1]*Department of Biostatistics and Health Data Science, University of Pittsburgh, Pittsburgh, USA.*

[2]*Department of Biostatistics, University of Michigan, Ann Arbor, USA.*

ABSTRACT

Utilizing a novel entropy loss (EL) metric, a causal discovery method is proposed to understand directional effects in the causal relationship between depression and anxiety among medical interns. This method advances existing methods of bivariate causal discovery with theoretical guarantees of causal effect identifiability and statistical inference and enjoys good computational performance. Using data from the intern health study (n=6,858), the proposed method reveals with high statistical confidence that depression scores (PHQ-9) consistently predispose anxiety scores (GAD-7) across four longitudinal visits of the study, controlling for demographic confounders. This finding provides crucial insights into the directional effect useful for mental health intervention strategies for medical interns. Simulation studies demonstrate that EL achieves nearly superior accuracy compared to existing approaches across various conditions with reduced computation time. The EL framework's ability to handle discrete clinical scores while adjusting for confounders makes it particularly valuable for psychiatric epidemiology and broader applications in causal discovery with discrete data.

Keywords: Information theory; Asymmetry; Causal discovery; Mental Health

Statistical Methods for Chemical Mixtures: A Roadmap for Practitioners Using Simulation Studies and a Sample Data Analysis in the PROTECT Cohort

Wei Hao[1], Amber L. Cathey[2], Max M. Aung[3], Jonathan Boss[1], John D. Meeker[2], Bhramar Mukherjee[4]

[1]*Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA*

[2]*Department of Environmental Health Sciences, University of Michigan, Ann Arbor, MI, USA*

[3]*Division of Environmental Health, University of South California, Los Angeles, CA, USA*

[4]*Yale School of Public Health, Yale University, New Haven, CT, USA*

ABTRACT

Quantitative characterization of the health impacts associated with exposure to chemical mixtures has received considerable attention in current environmental and epidemiological studies. With many existing statistical methods and emerging approaches, it is important for practitioners to understand which method is best suited for their inferential goals. The goal of this paper is to provide empirical simulation-based evidence regarding performance of mixture methods to help guide researchers on selecting the best available methods to address three scientific questions in mixtures analysis: identifying important components of a mixture, identifying interactions among mixture components and creating a summary score for risk stratification and prediction. We conduct a review and comparison of 11 analytical methods available for use in mixtures research, through extensive simulation studies for continuous and binary outcomes. In addition, we carry out an illustrative data analysis using the PROTECT birth cohort from Puerto Rico, to examine the associations between exposure to chemical mixtures—metals, polycyclic aromatic hydrocarbons (PAHs), phthalates and phenols—and birth outcomes. Our simulation results suggest that the choice of methods depends on the goal of analysis and there is no clear winner across the board. For selection of important toxicants in the mixtures and for identifying interactions, Elastic net by Zou et al. (Enet), Lasso for Hierarchical Interactions by Bien et al. (HierNet), Selection of nonlinear interactions by a forward stepwise algorithm by Narisetty et al. (SNIF) have the most stable performance across simulation settings. For overall summary or a cumulative measure, we find that using the Super Learner to combine multiple Environmental Risk Scores can lead to improved risk stratification and prediction properties. We develop an integrated R package “CompMix” that provides a platform for mixtures analysis where the practitioners can implement a pipeline that includes

several approaches for mixtures analysis. Our study offers guidelines for selecting appropriate statistical methods for addressing specific scientific questions related to mixtures research. We identify critical gaps where new and better methods are needed.

Keywords: Chemical mixtures, environmental risk score

Back to Sessions List

Supervised Fusion Learning of Physical Activity Features: Functional Frameworks and Longitudinal Analysis with L_0 Regularization

Dr. Margaret Banker

Northwestern University Feinberg School of Medicine

ABTRACT

Wearable devices are crucial in physical activity research because they provide continuous, real-time monitoring of various health metrics such as heart rate, physical activity, sleep patterns, and vital signs. These devices enable the collection of extensive, longitudinal data, offering insights into the daily lives and health trajectories of older adults. This information is invaluable for identifying early signs of health decline, assessing the effectiveness of interventions, and personalizing care plans. I consider wearable device data in a functional framework with an L_0 regularization approach, handling highly correlated micro-activity windows that serve as predictors in a scalar-on-function regression model. I develop a longitudinal functional framework with repeated wearable data to understand the influence of serially measured functional accelerometer data on longitudinal health outcomes. This method leverages Quadratic Inference Function (QIF) via mixed integer optimization for longitudinal data analysis to detect critical physical activity windows and assess their population-average effects on health outcomes.

Keywords: Actigraphy; Change-point detection; Mixed Integer Optimization (MIO); Scalar-on-function regression; Quadratic Inference Function (QIF)

Estimation and Inference of Quantile Spatially Varying Coefficient Models Over Complicated Domains

Myungjin Kim[1], Lily Wang[2,*], Huixia Judy Wang[3]

[1]*Department of Statistics, Kyungpook National University*

[2]*Department of Statistics, George Mason University*

[3]*Department of Statistics, Rice University*

ABSTRACT

This work presents a flexible quantile spatially varying coefficient model (QSVCM) for the regression analysis of spatial data. The proposed model enables researchers to assess the dependence of conditional quantiles of the response variable on covariates while accounting for spatial non-stationarity. Our approach facilitates learning and interpreting heterogeneity in spatial data distributed over complex or irregular domains. We introduce a quantile regression method that utilizes bivariate penalized splines in triangulation to estimate unknown functional coefficients. We establish the L2 convergence of the proposed estimators, demonstrating their optimal convergence rate under certain regularity conditions. An efficient optimization algorithm is developed using the alternating direction method of multipliers (ADMM). We develop wild bootstrap-based pointwise confidence intervals for the QSVCM quantile coefficients. Furthermore, we construct reliable conformal prediction intervals for the response variable using the proposed QSVCM. Numerical studies show the remarkable performance of the proposed methods.

Keywords: Alternating direction method of multiplier; Bivariate penalized spline; Nonparametric quantile regression; Triangulation

Leveraging External Individualized Prediction Models in Bayesian Survival Analysis

Mi-Ok Kim[1], Yena Jeon[1], Yunxiang Huang[1], Hangjoon Kim[2],

[1]*Department of Epidemiology and Biostatistics, University of California San Francisco, USA*

[2]*Division of Statistics and Data Science, University of Cincinnati, USA*

ABSTRACT

Individualized risk prediction algorithms, such as the Prostate Cancer Risk Assessment tool, are increasingly used to predict cancer relapse or progression. Since these algorithms are typically trained on large datasets, effectively integrating their outputs can enhance the efficiency of analyzing individual studies. In this research, we consider Cox regression analysis for right-censored time-to-event outcomes, incorporating external information provided by large-scale prediction models. We adopt a Bayesian inference in estimating the baseline hazard at each distinct time point. External information is integrated through the Kullback–Leibler (KL) divergence, leading to informative priors for Bayesian analysis. The performance of the proposed model is demonstrated through simulation studies and an application to data from a prostate cancer clinical trial.

Keywords: Informative Priors; External Prediction Information

Propensity Weighting Plus Adjustment in Proportional Hazards Model Is Not Doubly Robust

Erin Evelyn Gabriel*[1], Michael C Sachs[1] , Ingeborg Waernbaum[2] , Els Goetghebeur[3] , Paul F Blanche[1] , Stijn Vansteelandt[3] , Arvid Sjölander[4] , Thomas Scheike[1]

[1]*Section of Biostatistics, Department of Public Health, University of Copenhagen, København 1353, Denmark*

[2]*Department of Statistics, Uppsala University, Uppsala 75120, Sweden*

[3]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Gent 9000, Belgium

[4]*Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm 17177, Sweden*

ABSTRACT

Recently, it has become common for applied works to combine commonly used survival analysis modeling methods, such as the multivariable Cox model and propensity score weighting, with the intention of forming a doubly robust estimator of an exposure effect hazard ratio that is unbiased in large samples when either the Cox model or the propensity score model is correctly specified. This combination does not, in general, produce a doubly robust estimator, even after regression standardization, when there is truly a causal effect. We demonstrate via simulation this lack of double robustness for the semiparametric Cox model, the Weibull proportional hazards model, and a simple proportional hazards flexible parametric model, with both the latter models fit via maximum likelihood. We provide a novel proof that the combination of propensity score weighting and a proportional hazards survival model, fit either via full or partial likelihood, is consistent under the null of no causal effect of the exposure on the outcome under particular censoring mechanisms if either the propensity score or the outcome model is correctly specified and contains all confounders. Given our results suggesting that double robustness only exists under the null, we outline 2 simple alternative estimators that are doubly robust for the survival difference at a given time point (in the above sense), provided the censoring mechanism can be correctly modeled, and one doubly robust method of estimation for the full survival curve. We provide R code to use these estimators for estimation and inference in the supporting information.

Keywords: causal inference; Cox model; double robustness; inverse probability of treatment weighting; parametric proportional hazards

Event History Regression with Pseudo-Observations: Computational Approaches and Causal Inference

Michael C Sachs[1], Erin E Gabriel[1]

Section of Biostatistics and the Pioneer Centre for SMARTbiomed, University of Copenhagen

ABSTRACT

Due to tradition and ease of estimation, the vast majority of clinical and epidemiological papers with time-to-event data report hazard ratios from Cox proportional hazards regression models. Although hazard ratios are well known, they can be difficult to interpret, particularly as causal contrasts, in many settings. Nonparametric or fully parametric estimators allow for the direct estimation of more easily causally interpretable estimands such as the cumulative incidence and restricted mean survival. However, modeling these quantities as functions of covariates is limited to a few categorical covariates with nonparametric estimators, and often requires simulation or numeric integration with parametric estimators. Combining pseudo-observations based on non-parametric estimands with parametric regression on the pseudo-observations allows for the best of these two approaches and has many nice properties. In this talk, we review a user friendly, easy to understand way of doing event history regression for the cumulative incidence and the restricted mean survival, using the pseudo-observation framework for estimation. The method uses the well known formulation of a generalized linear model and allows for extensions to double-robust estimation.

Keywords: survival analysis; competing risks; pseudo observations; regression; causal inference

Transporting Evidence from and to External Studies by Leveraging Aggregate Data

Ying Sheng[1], Yifei Sun[2], **Chiung-Yu Huang[3]**

[1]*State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China*

[2]*Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, USA*

[3]*Department of Epidemiology & Biostatistics, University of California at San Francisco, California, USA*

ABSTRACT

With the increasing availability of data in the public domain, there has been a growing interest in exploiting information from various sources to facilitate the decision-making processes. However, in real-world applications, particularly those dealing with sensitive areas such as healthcare and finance, individual-level data are often unavailable, leaving only aggregate data from external sources. In this talk, I will demonstrate how one can leverage the external aggregate data to improve the estimation efficiency in smaller-scale studies via the empirical likelihood framework. This approach can accommodate the heterogeneity and uncertainty in external information simultaneously. Under a similar framework, I will also introduce a novel approach for transporting evidence from clinical studies to target populations using only covariate summary statistics to account for distributional shifts and uncertainty in the external information. Conditions to ensure the validity of the proposed estimators will be examined.

Keywords: Covariate shift; Entropy balancing; Exponential tilting; Meta-Analysis; Transportability

Data Secure Transfer Learning from Heterogeneous Low Rank and Sparse Panel VAR Models

George Michailidis

University of California, Los Angeles

ABSTRACT

Transfer learning (TL) enhances high-dimensional inference by borrowing strength from auxiliary data sets. We extend TL to multivariate time series settings through a panel Vector Autoregression (VAR) model, whose coefficients decompose into heterogeneous low-rank and sparse components. The proposed algorithm treats the two components independently and automatically detects the transferable auxiliary sources for each. Importantly, it operates solely on *pre-trained estimators* from the *auxiliary* panel, thereby preserving data privacy. We establish that the source detection mechanism achieves high accuracy under mild technical conditions. The proposed TL approach delivers significant performance improvements in data-scarce regimes---precisely the scenarios that motivate transfer learning. Extensive numerical experiments based on synthetic data, conducted for both correctly specified and misspecified time series models, along with applications to macroeconomic time series, confirm the practical effectiveness of our method. The results highlight the potential of TL to enhance predictive tasks in complex multivariate time series models when primary data are limited.

Fast Segmentation of Watermarked Texts from Large Language Models through Epidemic Change-Points Framework

Soham Bonnerjee, Subhrajyoty Roy, **Sayar Karmakar**

University of Florida

ABSTRACT

With the increasing popularity of large language models, concerns over content authenticity have led to the development of myriad watermarking schemes. These schemes can be used to detect a machine-generated text via an appropriate key, while being imperceptible to readers with no such keys. The corresponding detection mechanisms usually take the form of statistical hypothesis testing for the existence of watermarks, spurring extensive research in this direction. However, the finer-grained problem of identifying which segments of a mixed-source text are actually watermarked, is much less explored; the existing approaches either lack scalability or theoretical guarantees robust to paraphrase and post-editing. In this work, we introduce a unique perspective to such watermark segmentation problems through the lens of epidemic change-points. By highlighting the similarities as well as differences of these two problems, we motivate and propose WISER: a novel, computationally efficient, watermark segmentation algorithm. We theoretically validate our algorithm by deriving finite sample error-bounds, and establishing its consistency in detecting multiple watermarked segments in a single text. Complementing these theoretical results, our extensive numerical experiments show that WISER outperforms state-of-the-art baseline methods, both in terms of computational speed as well as accuracy, on various benchmark datasets embedded with diverse watermarking schemes. Our theoretical and empirical findings establish WISER as an effective tool for watermark localization in most settings. It also shows how insights from a classical statistical problem can lead to a theoretically valid and computationally efficient solution of a modern and pertinent problem.

Keywords: Epidemic change, Large language model, Anomaly detection

Monitoring and Early Detection of Instability in Manufacturing Process Using Vector Autoregression Models

Yi-Ting Wang[1], Ying-Chao Hung[2]

Institute of Industrial Engineering, National Taiwan University

ABSTRACT

Instability in manufacturing processes (e.g., semiconductor wafer fabrication) can cause severe economic losses if not detected early. This talk presents a monitoring framework based on Vector Autoregression (VAR) models to detect and predict whether a process is drifting out of control. We first show how linear predictors can characterize stability and serve as monitoring statistics for online data streams. We then design monitoring strategies with rigorous control of the family-wise error rate (FWER), ensuring reliable inference in multivariate settings. An optimal strategy is introduced to balance early detection with false alarms by minimizing a weighted combination of FWER and Type II error. The framework also extends to cases with partial variable information and incorporates fault propagation and causal analysis under a Bayesian framework. Together, these methods form a practical toolkit for early detection and diagnosis of process instability, supporting more resilient and efficient manufacturing operations.

Keywords: Vector Autoregression; Process Monitoring; Fault Detection; Family-wise ErrorRate; Bayesian Causal Analysis

Recent Advances of Structural Biology via Single Particle Cryogenic Electron Microscopy and the Remaining Challenges

Wei-Hau Chang
Institute of Chemistry, Academia Sinica

ABSTRACT

Single-particle cryogenic electron microscopy (cryo-EM) has revolutionized molecular studies in life science by enabling high-resolution structure determination of biological macromolecules without requiring crystallization. This revolution was propelled by the advent of direct electron detectors, advanced algorithms, and modern cryo-EM with increased high-throughput data collection schemes, where routine acquisition of a large dataset containing sub-millions to millions of particle images can be achieved in a single day. Compared to X-ray crystallography, this technique has unique strength in simultaneous capturing of a mixture of solution conformations reflecting functional states to aid the mechanistic understanding of dynamic biological process. However, as single cryo-EM imaging is performed with low-dose electron exposure (10-40 $e^-/Å^2$) to minimize radiation damage, the produced images are highly noisy with signal-to-noise ratios (SNRs) typically approaching ~0.1 or lower. As a result, reconstructing accurate 3D structures from such data presents a challenging inverse problem, for which the solution entails a multi-step workflow of image analysis that progressively harnesses the weak signals from the noises. In this lecture, I will first briefly explain the workflow including the major bottle met by 2D classification for particle image curation and other challenges. In addition, I will gives a number of successful examples of structure determination using this approach from our laboratories to illustrate how this method has been utilized to solve issues ranging from greenhouse gas to fish farming.

CRISP: A Modular Platform for Cryo-EM Image Segmentation and Processing with Conditional Random Field

Szu-Chi Chung[1*], Po-Cheng Chou[1]

Department of Applied Mathematics, National Sun Yat-sen University

ABSTRACT

Distinguishing signal from background in cryogenic electron microscopy (cryo-EM) micrographs is a critical yet challenging step due to the inherently low signal-to-noise ratio (SNR), the presence of contaminants, variable ice thickness, and densely packed particles of heterogeneous sizes. Recent image segmentation methods, which operate at the pixel level, offer several advantages over traditional object-detection approaches: they enable more accurate suppression of false positives via segmented blob mass, improve particle centering by leveraging full brightness profiles, and more reliably detect irregularly shaped particles. However, the low SNR in cryo-EM data makes it difficult to obtain accurate pixel-level annotations for training, and in the absence of standardized evaluation platforms, most existing segmentation pipelines rely on ad hoc design choices.

Here, we present a modular platform for generating high-quality reference segmentation maps automatically. The platform supports flexible combinations of segmentation architectures, feature extractors, and loss functions, and integrates a novel Conditional Random Fields (CRF) module that uses class-discriminative features and a regularized optimization scheme to refine coarse predictions into fine-grained masks. On synthetic datasets, models trained with our reference labels achieve over 90% accuracy, precision, recall, Intersection-over-Union (IoU), and F1 score at the pixel level. Furthermore, we demonstrate that the predicted segmentation maps can be directly used for particle picking, yielding higher-resolution 3D density maps from real cryo-EM datasets — matching expert-curated reconstructions and surpassing the performance of state-of-the-art particle-picking tools.

Keywords: cryogenic electron microscopy; image segmentation; image processing, conditional random fields

A Robust Hierarchical Linear Model for Cryo-EM Analysis

I-Ping Tu

Institute of Statistical Science, Academia Sinica

ABSTRACT

Cryo-electron microscopy (cryo-EM) is essential for determining high-resolution three-dimensional maps of biological macromolecules, which subsequently facilitate atomic model generation. The atomic models and their corresponding cryo-EM maps are archived in the Protein Data Bank (PDB) and the Electron Microscopy Data Bank (EMDB), respectively. In this study, we present a robust hierarchical modeling approach that quantitatively links cryo-EM maps with atomic models. Our model accounts for parameter variations based on atom types and amino acid categories, effectively capturing the nuanced influences of the local electron distribution environment. By applying the minimum power divergence method, we derive robust estimates that minimize the impact of outliers. The effectiveness and robustness of our approach have been validated through applications to both simulated and real datasets.

Wen Zhou

Factor Models of Matrix-Valued Time Series: Nonstationarity and Cointegration

Degui Li, Yayi Yan, Qiwei Yao

University of Macau, Shanghai University of Finance and Economics, London School of Economics

ABSTRACT

In this paper, we consider the nonstationary matrix-valued time series with common stochastic trends. Unlike the traditional factor analysis which flattens matrix observations into vectors, we adopt a matrix factor model in order to fully explore the intrinsic matrix structure in the data, allowing interaction between the row and column stochastic trends, and subsequently improving the estimation convergence. It also reduces the computation complexity in estimation. The main estimation methodology is built on the eigenanalysis of sample row and column covariance matrices when the nonstationary matrix factors are of full rank and the idiosyncratic components are temporally stationary, and is further extended to tackle a more flexible setting when the matrix factors are cointegrated and the idiosyncratic components may be nonstationary. Under some mild conditions which allow the existence of weak factors, we derive the convergence theory for the estimated factor loading matrices and nonstationary factor matrices. In particular, the developed methodology and theory are applicable to the general case of heterogeneous strengths over weak factors. An easy-to-implement ratio criterion is adopted to consistently estimate the size of latent factor matrix. Both simulation and empirical studies are conducted to examine the numerical performance of the developed model and methodology in finite samples.

Keywords: Common stochastic trends, Eigenanalysis, Matrix error-correction models, Matrix factor models, Ratio criterion

Estimating SNR in High-Dimensional Linear Models

Xiaodong Li

Department of Statistics, University of California, Davis

ABSTRACT

This talk develops robust methods for estimating signal-to-noise ratios (SNR) and variance components in high-dimensional linear models. We first show that the random-effects MLE remains consistent and asymptotically normal under substantial model misspecification, including fixed coefficients and heteroskedastic errors. We then extend the method-of-moments framework to multivariate responses, deriving asymptotic distributions using moment identities of the Wishart distribution. The resulting procedures require no sparsity assumptions and provide heteroskedasticity-robust inference through an explicit variance–inflation correction. Simulations demonstrate that the proposed confidence intervals achieve reliable coverage across a wide range of high-dimensional settings.

Keywords: high-dimensional inference; random effects; signal-to-noise ratio.

Natural Covariate-Adjusted Graphical Regression

Ruobin Liu[1], **Guo Yu[1]**

[1]*Department of Statistics and Applied Probability, University of California Santa Barbara*

ABSTRACT

Gaussian graphical models (GGMs) are widely used for recovering the conditional independence structure among random variables. Recently, several key advances have been made to exploit an additional set of variables for better estimating the GGMs of the variables of interest. For example, in co-expression quantitative trait locus (eQTL) studies, both the mean expression level of genes as well as their pairwise conditional independence structure may be adjusted by genetic variants local to those genes. Existing methods to estimate covariate-adjusted GGMs either allow only the mean to depend on covariates or suffer from poor scaling assumptions due to the inherent non-convexity of simultaneously estimating the mean and precision matrix. In this paper, we propose a convex formulation that jointly estimates the covariate-adjusted mean and precision matrix by utilizing the natural parametrization of the multivariate Gaussian likelihood. This convexity yields theoretically better performance as the sparsity and dimension of the covariates grow large relative to the number of samples. We verify our theoretical results with numerical simulations and perform a reanalysis of an eQTL study of glioblastoma multiforme (GBM), an aggressive form of brain cancer.

Keywords: Gaussian Graphical Models; Graphical Regressions; Convex Parameterization

The Blurred Line between Genes and Environments: Insights from GWAS of Family Members' Phenotypes

Qiongshi Lu[1]

[1]*Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison*

ABSTRACT

Genome-wide association study (GWAS) methodologies have become quite standard for complex trait genetic research. Today, a modern GWAS typically correlates a phenotype with tens of millions of genetic variants in large cohorts of millions of individuals to reveal genotype-phenotype associations. However, this seemingly standard approach can give largely biased and/or confounded results in various applications. In this talk, I will discuss a new study design which associates genetic data of a cohort with their family members' phenotypes. That is, the genotypic and phenotypic variables in the GWAS are collected from different individuals. Through three separate applications, focusing on offspring, parental, and spousal phenotypes, I will discuss several challenges and new insights in genetic nurture, ascertainment bias, and assortative mating. The phenotypes discussed in this talk will include socioeconomic outcomes, neurodegenerative disease risk, as well as human partner choice.

Keywords: GWAS; within-family genomic analysis; assortative mating; social genomics.

Inferring Cell-Type-Specific Co-Methylation Networks from Single-Cell DNA Methylation Data

Kangyi Zhao[1], Wenzhuo Lin[2], **Jiebiao Wang[2,*]**, Rebecca Deek[2,*]

[1]*Department of Statistics, University of Pittsburgh*

[2]*Department of Biostatistics, University of Pittsburgh*

ABSTRACT

The recent development of single-cell DNA methylation (scDNAm) technologies has enabled the detailed collection of cell-type-specific epigenetic regulation data. Yet, co-methylation network estimation and inference remain challenging due to data sparsity, zero-one inflation, and complex dependencies. Here we present scCoNet, a novel statistical framework for cell-type-specific co-methylation inference using a copula model with zero-one-inflated beta marginal distributions. Our approach flexibly captures both the marginal and joint methylation distributions while accounting for sparsity and leveraging copulas for dependence modeling. We identify distinct epigenetic modules and regulatory pathways by integrating cell type information and covariates. We demonstrate the utility of our method on simulated and real scDNAm data, uncovering biologically meaningful co-methylation patterns linked to cell functions. This work offers a powerful tool to decipher the epigenetic landscape at single-cell resolution and illuminates cell-type-specific regulatory mechanisms.

Keywords: Statistical genomics; DNA methylation; single cell; network

Single-Cell Multiomic Analysis of Circadian Rhythmicity

Chun Yip Tong[1], Jerome Menet[1], **Yuchao Jiang[1,2]**

[1]*Department of Biology, College of Arts and Sciences, Texas A&M University.*

[2]*Department of Statistics, College of Arts and Sciences, Texas A&M University.*

ABSTRACT

Circadian rhythms are remarkably widespread across most organisms, regulating hormonal, metabolic, physiological, and behavioral oscillations through molecular clocks that orchestrate the rhythmic expression of thousands of genes. Here, we generate single-cell RNA and ATAC multiomics data to simultaneously characterize gene expression and chromatin accessibility of mouse liver cells across the 24-hour day. We interrogate multimodal circadian rhythmicity in both discretized cell types and transient sub-lobule cell states, capturing space-time omics profiles. We delve beyond mean cyclic patterns to characterize stochastic transcriptional bursting and infer spatiotemporal gene regulatory networks that control circadian rhythmicity and liver physiology. Our findings apply to existing single-cell data of mouse and Drosophila brains and are validated by time-series single-molecule fluorescence in situ hybridization and vast amounts of orthogonal omics data. Altogether, our study constructs a comprehensive map of the time-series transcriptomic and epigenomic landscapes that elucidate the function and mechanism of the liver peripheral clocks.

Keywords: circadian rhythm; single-cell multiomics; RNA expression and DNA accessibility; transcriptional regulation

mist: A Hierarchical Bayesian Framework for Detecting Differential DNA Methylation Dynamics in Single-Cell Data

Hao Feng

Department of Biostatistics and Data Science,

The University of Texas Health Science Center at Houston

ABSTRACT

Recent advancements in single-cell DNA methylation (scDNAm) sequencing technologies have enabled the profiling of epigenetic landscapes at unprecedented resolution, offering insights into cellular heterogeneity, differentiation and evolution. Trajectory inference, which orders cells along pseudotime, allows researchers to track genomics changes across continuous cell states and identify key loci exhibiting differential methylation. However, no methods currently exist to model methylation changes along pseudotime in scDNAm data. Here, we present a hierarchical Bayesian framework for scDNAm data analysis. Our method, named mist (methylation inference for single-cell along trajectory), models stage-specific biological variations, identifies genomic features with significant methylation changes along pseudotime, and performs Differential Methylation (DM) analysis across phenotypical groups. Simulations demonstrate its superior accuracy in detecting DM genes along pseudotime compared to existing methods. Applied to multi-omics datasets of mouse embryonic development and developing human brain, mist identifies key developmental regulators, whose methylation patterns align with lineage transitions. mist is publicly available as an R/Bioconductor package at https://bioconductor.org/packages/mist.

Keywords: epigenetics; bioinformatics; computational biology; biostatistics

CQUESST: A Bayesian Framework for Soil-Carbon Sequestration

Noel Cressie[1], Dan Pagendam[2]

[1]-*School of Mathematics and Applied Statistics, University of Wollongong, Australia*

[2]-*Data 61, CSIRO, Australia*

ABSTRACT

A statistical framework we call CQUESST (Carbon Quantification and Uncertainty from Evolutionary Soil STochastics), which models carbon sequestration and cycling in soils, is applied to a long-running agricultural experiment that controls for crop type, tillage, and season. CQUESST embeds a dynamic stochastic model of soil carbon, motivated by the deterministic RothC soil-carbon model, within a Bayesian hierarchical statistical model. CQUESST has a coherent framework that acknowledges uncertainties in soil-carbon dynamics, in physical parameters, and in observations. The long-running experiment ran from 2000-2010 and is called the Millenium Tillage Trial; here CQUESST is used to model soil carbon in six pools, across 42 agricultural plots, and on a monthly time-step for 10 years. It is implemented efficiently in the probabilistic programming language Stan using its MapReduce parallelisation. We infer the effectiveness of different experimental treatments for soil-carbon sequestration; and we show how CQUESST can be used for the analysis of designed experiments to draw statistically defensible conclusions about the dependence of soil-carbon decay rates on crop rotations and tillage treatments. These results take into account the uncertainties in the model, resulting in inferences that could inform soil-carbon-sequestration decisions and policies. This joint research is also with Jeff Baldock, David Clifford, Ryan Farquharson, Lawrence Murray, and Mike Beare.

Keywords: Bayesian hierarchical statistical model; biophysical-statistical modelling; posterior inference; soil-carbon cycling

Modelling Count Data in the Presence of Intervention

Abdel H. El-Shaarawi

Department of Statistics, Cairo University, Egypt

ABSTRACT

Analysis of count data is extremely important in environmental risk assessment particularly when developing control limits on pollutants entering the environment from point and non-point sources that harm the resident biota. Poisson and mixed Poisson models and their extensions provide the backbone of inferential methods for dealing with various complexity encountered in the analysis of count data. Here we show how to adopt these methods to deal with over and under dispersion and lack of independence in real data sets from the Canadian Effects Monitoring Program on aquatic biota, the drinking water project of the International Development Research Center to help developing countries to have safe drinking water; The ELF Ecological Monitoring Program of the U.S Navy and the 2024 Canadian wildfire on the city of Jasper.

Keywords: Toxic contaminants, Impact Assessment, Poisson regression; negative binomial, lognormal, Taylo's power low

Covariate-Dependent Spatio-Temporal Covariance Models

Yen-Shiu Chin[1], **Nan-Jung Hsu[2]**, Hsin-Cheng Huang[1]

[1]*Institute of Statistical Science, Academia Sinica*

[2]*Institute of Statistics and Data Science, National Tsing-Hua University*

ABSTRACT

Geostatistical regression models are widely used in environmental and geophysical sciences to characterize the mean and dependence structures for spatio-temporal data. Traditionally, these models account for covariates solely in the mean structure, neglecting their potential impact on the spatio-temporal covariance structure. This paper addresses a significant gap in the literature by proposing a novel covariate-dependent covariance model within the spatio-temporal random effects model framework. Our approach integrates covariates into the covariance function through a Cholesky-type decomposition, ensuring compliance with the positive-definite condition. We employ maximum likelihood for parameter estimation, complemented by an efficient expectation conditional maximization algorithm. Simulation studies demonstrate the superior performance of our method compared to conventional techniques that ignore covariates in spatial covariance. We further apply our model to a PM2.5 dataset from Taiwan, highlighting wind speed's pivotal role in influencing the spatio-temporal covariance structure. Additionally, we incorporate wind speed and sunshine duration into the covariance function for analysing Taiwan ozone data, revealing a more intricate relationship between covariance and these meteorological variables.

Keywords: Cholesky decomposition; Maximum likelihood; Nonstationary spatial covariance function; Vector autoregressive model

Model-Free Inference for Characterizing Protein Mutations through a Coevolutionary Lens

Fan Yang[1], **Zhao Ren[1]**, Wen Zhou[2], Robert Jernigan[3]

[1]*Department of Statistics, University of Pittsburgh,*

[2]*Department of Biostatistics, New York University,*

[3]*Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University*

ABSTRACT

Multiple sequence alignment (MSA) data play a crucial role in the study of protein mutations, with contact prediction being a notable application. Existing methods are often model-based or algorithmic and typically do not incorporate statistical inference to quantify the uncertainty of the prediction outcomes. To address this, we propose a novel framework that transforms the task of contact prediction into a statistical testing problem. Our approach is motivated by the partial correlation for continuous random variables. With one-hot encoding of MSA data, we are able to construct a partial correlation graph for multivariate categorical variables. In this framework, two connected nodes in the graph indicate that the corresponding positions on the protein form a contact. A new spectrum-based test statistic is introduced to test whether two positions are partially correlated. Moreover, the new framework enables the identification of amino acid combinations that contribute to the correlation within the identified contacts, an important but largely unexplored aspect of protein mutations. Numerical experiments demonstrate that our proposed method is valid in terms of controlling Type I errors and powerful in general. Real data applications on various protein families further validate the practical utility of our approach in coevolution and mutation analysis.

Keywords: Multivariate categorical data; Partial correlation; Precision matrix; Multiple sequence alignment; Protein mutation

Microbial Causal Mediation Analysis under Spatially Correlated Exposure

Sooran Kim, Chan Wang, Soyoung Kwak, Jiyoung Ahn, **Huilin Li**

New York University, Grossman School of Medicine, Department of Population Health

ABSTRACT

Recent research suggests that the environmental exposure has been linked with the human microbiome, which can contribute to human health and disease. In particular, air pollution exposure plays a critical role in cancer presentation and prognosis. Understanding the complex interplay among disease status, microbiome abundances, and environmental exposure is therefore essential. Environmental exposures, such as air pollution, are often spatially autocorrelated, posing significant challenges for statistical analysis. In this work, we explore the associations among three key factors—disease status, environmental exposures, and microbiome abundances—while incorporating the spatial dependence of environmental exposures and the intrinsic correlation within microbiome.

Key words: Microbiome, causal mediation, spatial correlation

Testing Composite Null Hypotheses with High-Dimensional Dependent Data: A Computationally Scalable FDR-Controlling Procedure

Pengfei Lyu, Xianyang Zhang and **Hongyuan Cao**

Duke University

Texas A & M University

Florida State University

ABSTRACT

Testing composite null hypotheses is fundamental to many scientific applications, including mediation and replicability analyses, and becomes particularly challenging in high-throughput settings involving tens of thousands of features. Existing high dimensional composite null hypotheses testing often ignores the dependence structure among features, leading to overly conservative or liberal results. To address this limitation, we develop a four-state hidden Markov model (HMM) for bivariate p-value sequences arising from two-study replicability analysis. This model captures local dependence among features and accommodates study-specific heterogeneity. Based on the HMM, we propose a multiple testing procedure that asymptotically controls the false discovery rate (FDR). Extending this framework to more than two studies is computationally intensive, with complexity growing exponentially in the number of studies $n.$ To address this scalability issue, we introduce a novel e-value framework that reduces computational complexity to quadratic in $n,$ while preserving asymptotic FDR control. Extensive simulations demonstrate that our method achieves higher power than existing approaches at comparable FDR levels. When applied to genome-wide association studies (GWAS), the proposed approach identifies novel biological findings that are missed by current methods.

Keywords: Composite null hypotheses, e-values, false discovery rate, hidden Markov model, high dimension, non-parametric maximum likelihood estimation

D-CDLF: Decomposition of Common and Distinctive Latent Factors for Multi-view High-Dimensional Data

Hai Shu[1], Hongtu Zhu[2]

[1]*Department of Biostatistics, New York University*

[2]*Department of Biostatistics, The University of North Carolina at Chapel Hill*

ABSTRACT

Modern biomedical studies often collect multi-view data, that is, multiple types of data measured on the same set of objects. A typical approach to the joint analysis of multiple high-dimensional data views is to decompose each view's data matrix into three parts: a low-rank common-source matrix generated by common latent factors of all data views, a low-rank distinctive-source matrix generated by distinctive latent factors of the corresponding data view, and an additive noise matrix. Existing decomposition methods often focus on the uncorrelatedness between the common latent factors and distinctive latent factors, but inadequately address the equally necessary uncorrelatedness between distinctive latent factors from different data views. We propose a novel decomposition method, called Decomposition of Common and Distinctive Latent Factors (D-CDLF), to effectively achieve both types of uncorrelatedness. Consistent estimators of our D-CDLF method are established, demonstrating reasonably good finite-sample numerical performance. The superiority of D-CDLF over state-of-the-art methods is corroborated by simulations and the analysis of imaging and genomic data from the Alzheimer's Disease Neuroimaging Initiative.

Keywords: Canonical correlation analysis; Common and distinctive latent factors; Data integration; Orthogonality constraint

Longitudinal First Hitting-Time Models with Extension to Neural Network

Mei-Ling Ting Lee

University of Maryland, College Park

ABSTRACT

Disease progression in a patient can be described mathematically as a stochastic process. The patient experiences a failure event when his/her disease progression first reaches a critical threshold level. This happening defines the failure event itself and the first hitting time (FHT) is the event time. First hitting-time based threshold regression (TR) models incorporate regression functions for parameters of the underlying stochastic process. The TR models are intuitive and do not require the proportional hazards assumption, therefore represent a realistic alternative to the Cox model. Recently the TR model has been extended to the Levy family with stationary independent increments and a cumulant generating function. Extension to neural network applications will also be discussed.

Keywords: Levy processes, non-proportional hazards, Wiener processes

Time-Dependent Pseudo R-Squared for Assessing Predictive Performance in Competing Risks Data

Zian Zhaung, Wen Su, Eric Kawaguchi, and **Gang Li***

Biostatistics and Computational Medicine, University of California at Los Angeles

ABSTRACT

Evaluating and validating the performance of prediction models is a fundamental task in statistics, machine learning, and their diverse applications. However, developing robust performance metrics for competing risks time-to-event data poses unique challenges. We first highlight how certain conventional predictive performance metrics for competing risks time-to-event data, such as the C-index, Brier Score, and time-dependent AUC, can yield unexpected results when comparing predictive performance between different prediction models. To address this research gap, we introduce a novel time-dependent pseudo R-squared measure to evaluate the predictive performance of a predictive cumulative incidence function over a restricted time domain under right-censored competing risks time-to-event data. Specifically, we first propose a population-level time-dependent pseudo R-squared measures for the competing risk event of interest and then define their corresponding sample versions based on right-censored competing risks time-to-event data. We investigate the asymptotic properties of the proposed measure and demonstrate its advantages over conventional metrics through comprehensive simulation studies and three real-data applications.

Keywords: Right censoring; Competing risks; Explained variance; Predictive performance

Assessing Transcriptomic Heterogeneity of Single-Cell RNASeq Data by Bulk-Level Gene Expression Data

Khong-Loon Tiong[1#], Dmytro Luzhbin[1#], **Chen-Hsiang Yeang[1*]**

[1]*Institute of Statistical Science, Academia Sinica, Taipei, Taiwan*

[#]*Co-first authors*

[*]*Corresponding author, chyeang@stat.sinica.edu.tw*

Abstract

Motivation

Single-cell RNA sequencing (sc-RNASeq) data illuminate transcriptomic heterogeneity but also possess a high level of noise, abundant missing entries and sometimes inadequate or no cell type annotations at all. Bulk-level gene expression data lack direct information of cell population composition but are more robust and complete and often better annotated. We propose a modeling framework to integrate bulk-level and single-cell RNASeq data to address the deficiencies and leverage the mutual strengths of each type of data and enable a more comprehensive inference of their transcriptomic heterogeneity. Contrary to the standard approaches of factorizing the bulk-level data with one algorithm and (for some methods) treating single-cell RNASeq data as references to decompose bulk-level data, we employed multiple deconvolution algorithms to factorize the bulk-level data, constructed the probabilistic graphical models of cell-level gene expressions from the decomposition outcomes, and compared the log-likelihood scores of these models in single-cell data. We term this framework *backward deconvolution* as inference operates from coarse-grained bulk-level data to fine-grained single-cell data. As the abundant missing entries in sc-RNASeq data have a significant effect on log-likelihood scores, we also developed a criterion for inclusion or exclusion of zero entries in log-likelihood score computation.

Results

We selected six deconvolution algorithms and validated backward deconvolution in four datasets. In the insilico mixtures of mouse sc-RNASeq data, the log-likelihood scores of the deconvolution algorithms were strongly anticorrelated with their errors of mixture coefficients and cell type specific gene expression signatures. In the true bulk-level mouse data, the sample mixture coefficients were unknown but the loglikelihood scores were strongly correlated with accuracy rates of inferred cell types. In datasets of breast cancer and low-grade gliomas (LGG), we compared the log-likelihood scores of three simple hypotheses about the gene expression patterns of the cell types underlying the tumor subtypes. The model that tumors of each subtype were dominated by one cell type persistently outperformed an alternative model that each cell

type had elevated expression in one gene group and tumors were mixtures of those cell types. The results indicate that backward deconvolution serves as a sensible model selection tool for deconvolution algorithms and facilitates discerning hypotheses about cell type compositions underlying heterogeneous specimens such as tumors.

Availability and implementation

We have implemented the backward deconvolution algorithm in R, and deposited the source codes and their description on github.com/chyeang/backward-deconvolution/.

Keywords: Single-cell RNASeq data, Deconvolution, Probabilistic graphical models, Heterogeneity

Back to Sessions List

A Unified Framework for Statistical Inference and Power Analysis of Single and Comparative F_β Scores

Chih-Yuan Hsu, Qi Liu, Yu Shyr

Department of Biostatistics, Vanderbilt University Medical Center, Nashville 37203, TN, USA

ABSTRACT

Machine learning and artificial intelligence are increasingly applied to medical diagnostics and clinical decision-making. To evaluate model performance, the F1 score and its generalized form, the F score, are widely used as they balance precision and sensitivity. However, rigorous statistical inference and power analysis for the F1 and F scores remain limited. In this study, we propose psF1, a unified and comprehensive framework for interval estimation, hypothesis testing, and power and sample size calculation for both single and comparative F1 and F scores. psF1 leverages exact probability distributions as well as normal approximations for large sample sizes to provide valid statistical inference and power analyses. Extensive simulations demonstrate the accuracy and robustness of psF1 across a range of sensitivity, precision, and sample size scenarios. We further showcase its practical utility through real-world biomedical classification tasks. This framework enables principled evaluation and comparison of classifiers using F1 and F scores with reliable uncertainty quantification and informed sample size planning. psF1 is freely available at http://github.com/cyhsuTN/psF1.

Keywords: F1 score; F score; Interval estimation; Hypothesis testing; Power and sample size calculation

Design Strategies for Robust Subsampling under Model Misspecification

Carlos de la Calle Arroyo[1], Laura Deldossi[2], Chiara Tommasi[3]

[1]*Department of Statistics and Operative Research, Universidad de Oviedo*

[2]*Department of Statistical Science, Università Cattolica del Sacro Cuore*

[3]*Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano*

ABSTRACT

Subsampling has become an increasingly important topic within experimental design, motivated by the growing computational burden of processing large-scale data and the high cost of data labeling in domains such as speech recognition. To address these challenges, a wide variety of subsampling methods have emerged, many of them extending principles of classical design theory to modern data-driven problems.

In this work, we focus on the problem of robust subsampling for prediction when the underlying response deviates from the linear model typically assumed by experimenters. Such model misspecification can severely compromise prediction accuracy, underscoring the need for design strategies that remain effective under deviations from idealized assumptions. We establish an explicit upper bound for the mean squared prediction error (MSPE) in this setting, providing theoretical insight into the predictive performance of subsampling schemes under model uncertainty.

Building on this result, we explore and compare several strategies for constructing efficient designs tailored to robust prediction. These strategies are evaluated against existing approaches, highlighting their advantages in terms of both theoretical guarantees and empirical performance.

Keywords: Design of Experiments; Subsampling; Prediction; I-Optimality

A Design Optimality Criterion Based on the AUC for Classification

Carlos de la Calle, **Jesus Lopez-Fidalgo**, Pablo Urruchi

Institute of Data Science and AI, University of Navarra

ABSTRACT

In binary classification, such as using the logistic model, optimal designs improve the estimation of the probabilities of being in one group after Maximum Likelihood fitting. Nevertheless, this could be not so good for the actual classification. This is the motivation for looking for an appropriate estimation method for classification and then an optimality criterion for searching for an optimal subsampling procedure. An example for classifying tweets coming either from humans or from bots is used to illustrate the results. The explanatory variables used are the open information given by Twitter: `friends', `listed', `favorites', `status' and `profile'. The last variable is binary while the rest are counts. Although this is not properly big data the interest of subsampling is that labeling the data is expensive.

Keywords: Classification, Area under the curve, Maximum likelihood, optimal design of experiments

Models and Procedures for the Estimation of Blood Alcohol Concentration in the Human Body

Juan M. Rodríguez-Díaz[1], Irene Mariñas-Collado[2], M. Teresa Santos-Martin[1]

[1]*Department of Statistics, University of Salamanca (Spain)*

[2]*Department of Statistics and Operations Research and Mathematics Didactics, University of Oviedo (Spain)*

ABSTRACT

The dynamics of alcohol elimination in the human body is very important in forensic and healthcare areas. Existing models often oversimplify with the assumption of linear elimination kinetics, limiting practical application, which should take into account the absorption phase as well. A simplified gamma model is propose to describe both periods, that can be adapted to avoid abrupt intakes and a certain lag in the alcohol absorption, as well as for considering multiple intakes.

In addition, the problem of estimating blood or breath alcohol concentration at a past time point will be addressed, which is a critical challenge in forensic science and legal medicine, especially when the time and amount of alcohol intake are uncertain. Retrocalculation methods are essential in cases like driving under the influence or drug-facilitated offenses, where delayed samples require precise modelling of alcohol kinetics. The work introduces a probabilistic framework that accounts for uncertainty in the time of ingestion, designing a procedure to assess the probability of exceeding legal thresholds at earlier times. These results provide a robust basis for hypothesis testing in forensic investigations, offering improved accuracy and interpretability under conditions of uncertainty.

Keywords: Blood Alcohol Concentration (BAC); Ethanol pharmacokinetics; Optimal Design of Experiments; Uncertainty Estimation

Swarm-Based Search Procedure for Finding Optimal Multi-Stage Designs for Phase II Clinical Trials

Ping-Yang Chen

Department of Statistics, National Taipei University

ABSTRACT

Multi-stage Phase II clinical trials offer advantages over single-stage designs by enabling interim analyses that can accurately inform early termination of the trial if there is evidence that the treatment is likely to be ineffective or effective. However, identifying optimal designs for multi-stage trials poses considerable computational challenges. In addition to having to optimize many integer-valued variables, there are multiple constraints, including order constraints. Traditional exhaustive search methods lack scalability and quickly become computationally infeasible when the number of stages is three or more. To overcome this challenge, we utilize a spherical coordinate system and reformulate the design problem as a continuous optimization task. The new formulation enables us to efficiently use Particle Swarm Optimization (PSO) to extend Simon's celebrated two-stage Phase II designs to three or more stages. Specifically, we show that our proposed search procedure not only reproduces the two-stage designs and certain three-stage designs found in the literature but also able to achieve the results more efficiently than traditional exhaustive search methods. We provide R codes for reproducing the optimal designs in this paper, which can be easily customized to generate tailor-made optimal designs for specific user needs.

Keyword: Adaptive Design, Maximum Sample Size, Minimax Design, Nature-Inspired Metaheuristics, Particle Swarm Optimization

Inference on Deep Neural Network Estimators

Yi LI

Department of Biostatistics, University of Michigan, Ann Arbor, USA

ABSTRACT

While deep neural networks (DNNs) are used for prediction, inference on DNN-estimated subject-specific means for categorical or exponential family outcomes remains underexplored. We address this by proposing a DNN estimator under generalized nonparametric regression models (GNRMs) and developing a rigorous inference framework. Unlike existing approaches that assume independence between estimation errors and inputs to establish the error bound, a condition often violated in GNRMs, we allow for dependence and our theoretical analysis demonstrates the feasibility of drawing inference under GNRMs. To implement inference, we consider an Ensemble Subsampling Method (ESM) that leverages U-statistics and the Hoeffding decomposition to construct reliable confidence intervals for DNN estimates. We show that, under GNRM settings, ESM enables model-free variance estimation and accounts for heterogeneity among individuals in the population.

Through simulations under nonparametric logistic, Poisson, and binomial regression models, we demonstrate the effectiveness and efficiency of our method. We further apply the method to the electronic Intensive Care Unit (eICU) dataset, a large scale collection of anonymized health records from ICU patients, to predict ICU readmission risk and offer patient-centric insights for clinical decision making.

Keywords: deep neural network; ensemble estimation; nonparametric regression.

Causal Learning with Label Noise: A Classification Approach for Paired Vectors

Grace Y. Yi*

University of Western Ontario

ABSTRACT

Causal inference involves determining whether a cause-effect relationship exists between two sets of interest, a task that can be framed as a binary classification problem. When dealing with a sequence of independent and identically distributed paired vectors, the kernel mean embedding of the probability distribution can be utilized to map the empirical distribution to a feature space. Subsequently, a classifier is trained in this feature space to predict causation for future vector pairs. However, this approach is susceptible to mislabeling of causal relationships, a common challenge in causation studies. In this talk, I will discuss the impact of mislabeled outputs on the training results. Moreover, I will present a learning method that takes into account the mislabeling effects and offer theoretical justifications for the validity of the proposed method.

Keywords: Binary classification, Causal inference, Label noise

In-Sample Evaluation of Subgroups Identified by Generic Machine Learning

Shuoxun Xu, **Xinzhou Guo***

Department of Mathematics, Hong Kong University of Science and Technology

ABSTRACT

When a subgroup is identified from the data, we must evaluate the post-hoc identified subgroup in a replicable way. The usual in-sample approach, which evaluates the post-hoc identified subgroup as predefined, might suffer from selection bias, and the issue can be exacerbated by generic machine learning-based subgroup identification and nonregularity; i.e. the boundary of the subgroup is non-smooth. The out-of-sample approach, which splits data into two parts—one for subgroup identification and the other for evaluation, can help address selection bias but might suffer from efficiency loss and instability issue, as the subgroup is identified using only part of the data. In this paper, we propose a conditional m-out-of-n perturbation approach to remove selection bias in in-sample subgroup evaluation and deliver valid inference on post-hoc identified subgroups when the subgroup is identified from the whole dataset of an observational study by generic machine learning. The proposed method is easy-to-compute and model-free, and remains valid regardless of whether regularity is satisfied. Through a novel theoretical framework of triple robustness linking rates of subgroups identification and nuisance estimation, we show that the proposed method, with an adaptive selection of the subsample size, achieves full efficiency across broad scenarios in generic machine learning for subgroup analysis. The merits of the proposed method are demonstrated by a re-analysis of the ACTG 175 trial.

Keywords: Asymptotically efficient; Conditional m-out-of-n perturbation; Model-free; Nonregularity; Selection bias; Triple robustness.

*The work was partially supported by a grant from Research Grants Council of the Hong Kong Special Administrative Region, China (HKUST 26308323), the Seed fund of the Big Data for Bio-Intelligence Laboratory (Z0428) and the grant L0438 from the Hong Kong University of Science and Technology

A Unified Framework of Analyzing Missing Data and Variable Selection Using Regularized Likelihood

Yuan Bian, Grace Yi, **Wenqing He***

Department of Statistical and Actuarial Sciences, University of Western Ontario

ABSTRACT

Missing data arise commonly in applications, and research on this topic has received extensive attention in the past few decades. Various inference methods have been developed under different missing data mechanisms, including missing at random and missing not at random. The assessment of a feasible missing data mechanism is, however, difficult due to the lack of validation data. The problem is further complicated by the presence of spurious variables in covariates. Focusing on missingness in the response variable, a unified modeling scheme is proposed by utilizing the parametric generalized additive model to characterize various types of missing data processes. Taking the generalized linear model to facilitate the dependence of the response on the associated covariates, the concurrent estimation and variable selection procedures are developed using regularized likelihood, and the asymptotic properties for the resultant estimators are rigorously established. The proposed methods are appealing in their flexibility and generality; they circumvent the need of assuming a particular missing data mechanism that is required by most available methods. Empirical studies demonstrate that the proposed methods result in satisfactory performance in finite sample settings. Extensions to accommodating missingness in both the response and covariates are also discussed.

Keywords: Missing data, Missing mechanism, Regularized likelihood

Logistics Regression Model for Differentially-Private Matrix Masking Data

Linh H. Nghiem[1], Aidong Adam Ding[2], **Samuel S. Wu[3]**

[1]*School of Mathematics and Statistics, University of Sydney, Sydney, Australia*

[2]*Department of Mathematics, Northeastern University, Boston, USA*

[3]*Health Informatics Institute, University of South Florida, Tampa, USA*

ABSTRACT

A recently proposed scheme utilizing local noise addition and matrix masking enables data collection while protecting individual privacy from all parties, including the central data manager. Statistical analysis of such privacy-preserved data is particularly challenging for nonlinear models like logistic regression. By leveraging a relationship between logistic regression and linear regression estimators, we propose the first valid statistical analysis method for logistic regression under this setting. Theoretical analysis of the proposed estimators confirmed its validity under an asymptotic framework with increasing noise magnitude to account for strict privacy requirements. Simulations and real data analyses demonstrate the superiority of the proposed estimators over naive logistic regression methods on privacy-preserved data sets.

Keywords: Logistic regression; Differential privacy; Matrix masking; Mixture normal

Partially-Global Fréchet Regression

Danielle C. Tucker and **Yichao Wu**[1]

Department of Mathematics, Statistics, and Computer Science, University of Illinois Chicago

ABSTRACT

We propose a partially-global Fréchet regression model by extending the profiling technique for the partially linear regression model (Severini and Wong 1992). This extension allows for the response to come from a generic metric space and can incorporate a combination of Euclidean predictors and a predictor which comes from another generic metric space. By melding together the local and global Fréchet regression models proposed by Petersen and Müller (2019), we gain a model that is more flexible than global Fréchet regression and more accurate than local Fréchet regression when the data generating process relies on a non-Euclidean predictor or is truly "global (linear)" for some scalar predictors. In this paper, we provide theoretical support for partially-global Fréchet regression and demonstrate its competitive finite-sample performance when applied to both simulated data and to real data which is too complex for traditional statistical methods.

Keywords: metric distance, partially linear model; random object

Learning nonparametric graphical model on heterogeneous network-linked data

Junhui Wang

Department of Statistics and Data Science The Chinese University of Hong Kong

ABSTRACT

Graphical models have been popularly used for capturing conditional independence structure in multivariate data, which are often built upon independent and identically distributed observations, limiting their applicability to complex datasets such as network-linked data. In this talk, we introduce a nonparametric graphical model that addresses these limitations by accommodating heterogeneous graph structures without imposing any specific distributional assumptions. The introduced estimation method effectively integrates network embedding with nonparametric graphical model estimation. It further transforms the graph learning task into solving a finitedimensional linear equation system by leveraging the properties of vectorvalued reproducing kernel Hilbert space. We will also discuss theoretical properties of the proposed method in terms of the estimation consistency and exact recovery of the heterogeneous graph structures. Its effectiveness is also demonstrated through a variety of simulated examples and a real application to the statistician coauthorship dataset.

Keywords: Graphical model, score matching, network-linked data

Residual-Based Subdata Selection for Local Linear Regression and Its Extension to Partial Linear Models

Chia-Wei Lin[1], Li-Shan Huang[2,*]

Institute of Statistics and Data Science, National Tsing Hua University

ABSTRACT

The rapid growth of data has introduced considerable computational challenges in statistical analysis. His study addresses this issue in local linear regression through representative subdata selection to reduce the computational burden, then extends the method to partial linear models. For local linear egression, a residual-based subdata selection (RESS) method is introduced. RESS yields a lower asymptotic mean squared error than existing methods in a neighborhood where the absolute asymptotic bias is largest. For partial linear models, an integrated method, termed IBRESS, combines ESS for the nonlinear component with information-based optimal subdata selection (IBOSS) for the linear component. IBRESS leverages the strengths of both methods and satisfies two theoretical properties: (i) similar to IBOSS, the convergence rate of the linear component depends on the full data size; and (ii) the nonlinear component retains the asymptotic properties of RESS. Simulation studies demonstrate that IBRESS reduces computational cost while maintaining estimation accuracy.

Keywords: semiparametric regression, big data, data reduction.

A Perturbation Subsampling for Large Scale Data

Yujing Yao[1] and **Zhezhen Jin[2]**

[1]*Department of Neurology*

[2]*Department of Biostatistics, Columbia University, New York, NY*

ABSTRACT

When analyzing large-scale data, subsampling methods and divide-and-conquer procedures are appealing, because they ease the computational burden, while preserving the validity of inferences. Here, sampling may occur with or without replacement. In this paper, we propose a perturbation subsampling approach based on independent and identically distributed stochastic weights for analyzing large-scale data. We justify the method based on optimizing convex objective functions by establishing the asymptotic consistency and normality of the resulting estimators. This method simultaneously provides consistent point and variance estimators. We demonstrate the finite-sample performance of the proposed method using simulation studies and two real-data analyses.

Keywords: Convex objective function, distributed computing, optimization, perturbation, subsampling.

Subgroup Mixture Challenges in Bridging Studies for Predictive Biomarkers

Szu-Yu Tang

Pfizer, Inc.

ABSTRACT

Predictive biomarkers play a critical role in precision medicine by identifying patient subgroups most likely to benefit from specific treatments. In scenarios where a companion diagnostic (CDx) is unavailable, patients are enrolled using a clinical trial assay (CTA) and subsequently retested with the CDx. This requires a bridging study to evaluate the clinical utility of the CDx, particularly its efficacy.

Recent insights from the Oncology Working Group (Liu, 2023) highlight potential logical inconsistencies arising from improper mixing of biomarker-positive and -negative subgroups in different efficacy endpoints. This research investigates the impact of subgroup mixture in the bridging study context. Through illustrative examples and simulation studies of different predictive and prognostic biomarker scenarios, this presentation will:

1. Examine the logical pitfalls associated with subgroup mixture in bridging studies.
2. Apply the Subgroup Mixture Estimation (SME) framework (Ding, 2016) to derive logically consistent estimates.
3. Examine the influence of assay concordance on subgroup mixture challenge.

These findings aim to enhance the robustness of CDx evaluation and support more reliable decision-making in precision oncology.

Keywords: bridging study, clinical trial assay (CTA), companion diagnostic test (CDx), subgroup mixture estimation (SME)

Scalable Joint Modeling of Multiple Longitudinal Biomarkers and Competing Risks Time-to-Event Data: with Applications to Mega-Scale Health Research

Shanpeng Li*[1,3], Emily Ouyang*[2], Jin Zhou[3], **Xinping Cui[2]**, Gang Li[3]

[1]*Department of Computational and Quantitative Medicine, City of Hope, Duarte, CA, USA.*

[2]*Department of Biostatistics, University of California at Riverside, Riverside, CA, USA.*

[3]*Department of Biostatistics, University of California at Los Angeles, Los Angeles, CA, USA.*

ABSTRACT

Joint modeling has become increasingly popular for characterizing the association between one or more longitudinal biomarkers and competing risks time-to-event outcomes. However, semiparametric multivariate joint modeling for large-scale data encounter substantial statistical and computational challenges, primarily due to the high dimensionality of random effects and the complexity of estimating nonparametric baseline hazards. These challenges often lead to prolonged computation time and excessive memory usage, limiting the utility of joint modeling for biobank-scale datasets. In this work, we introduce an efficient implementation of a semiparametric multivariate joint model, supported by a normal approximation and customized linear scan algorithms within an expectation-maximization (EM) framework. Our method significantly reduces computation time and memory consumption, enabling the analysis of data from thousands of subjects. The scalability and estimation accuracy of our approach are demonstrated through simulation studies. We also present an application to UK Biobank primary care study as an illustrative example. A user-friendly R package, FastJM, has been developed for the shared random effects joint model with efficient implementation. The package is publicly available on the Comprehensive R Archive Network https://CRAN.R-project.org/package=FastJM.

Keywords: longitudinal data, competing risks, normal approximation, linear scan algorithms, large-scale biobank data

Metaheuristics as a General-Purpose Optimization Tool for Statistical Research

Weng Kee Wong

Department of Biostatistics, Fielding School of Public Health

University of California at Los Angeles

ABSTRACT

Nature-metaheuristics have been widely used in engineering and computer science to address various types of optimization problems for decades and are now increasingly used across disciplines. They are increasingly popular in industry and academia for tackling all kinds of complex and high-dimensional optimization problems. Interestingly, metaheuristics seems to be still relatively underused in the statistical research community.

I present an overview of nature-inspired metaheuristics and some of their applications in statistics. The main appealing features of these algorithms are their speed, flexibility, availability of codes in different platforms, and ease of implementation and usage. Above all, they are virtually assumptions-free, which allows us to apply them to solve a huge range of optimization tasks. I will enumerate the advantages of nature-inspired metaheuristic algorithms over existing optimization algorithms and illustrate their diverse applications in biostatistics, and beyond. If time permits, I will demonstrate how nature-inspired algorithms can find more flexible and computationally challenging designs for early-phase clinical trials.

Keywords: Design Efficiency, Early Phase Trials, Optimal Experiment Designs, Particle Swarm Optimization. mixture regression models, multiple constraints, swarm intelligence

Mathematical Models of the Feedback between Population Dynamics and Biological or Cultural Evolution

Shota Shibasaki[1]

[1]*Faculty of Culture and Information Science, Doshisha University*

ABSTRACT

Temporal changes in populations are a central topic in biology. Ecologists have studied how the number of individuals within a population changes over time (i.e., population dynamics), while evolutionary biologists have examined how population-level characteristics, such as the mean and variance of trait values, evolve (i.e., evolutionary dynamics). Traditionally, evolutionary dynamics were considered much slower than population dynamics because evolution occurs through the accumulation of mutations and natural selection over many generations. However, over the past decade, both empirical and theoretical studies have shown that evolutionary dynamics can happen much faster than previously assumed; they can influence population dynamics and vice versa. Such rapid evolution, for example, allows a population to survive under environmental stresses that would otherwise lead to extinction. Moreover, recent research suggests that evolution occurs not only through genetic changes (biological evolution) but also through learning from others (cultural evolution). In this talk, I will first present a mathematical model and experimental results where a population can avoid extinction under environmental fluctuations by changing the mean and variance of trait values. I will then introduce a mathematical model of cultural evolution to explore the mechanisms for maintaining trait diversity within populations, which may influence population resilience and adaptability. Finally, I will outline an ongoing research direction: investigating how cultural evolution might influence population dynamics, and how its effects may differ from biological evolution.

Keywords: biological evolution; cultural evolution; mathematical model; population dynamics

Structure-Plasticity Interactions Shape Self-Organized Criticality in Neural Networks

Yoshiki A. Sugimoto[1], Masato S. Abe[2,3,4]

[1]*Graduate school of Culture and Information Science, Doshisha University*

[2]*Faculty of Culture and Information Science, Doshisha University, Kyoto, Japan*

[3]*Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan*

[4]*CBS-TOYOTA Collaboration Center, Saitama, Japan*

ABSTRACT

Multiple observations suggest that neural activity operates near the boundary between order and disorder—the vicinity of criticality—and this regime has been linked to broad dynamic range, efficient information transmission, and a balance between stability and flexibility. With self-organized criticality (SOC) in mind—namely, a feedback system that autonomously adjusts the system toward the critical regime through internal adaptive processes (e.g., synaptic plasticity) without relying on fine-tuned external parameters—this presentation adopts a framework that examines critical phenomena from two perspectives: (A) neuronal avalanches, in which activity-size distributions follow a power law, and (B) the edge of chaos, referring to flexible operating states near the boundary between chaotic and non-chaotic dynamics. From these viewpoints, we investigate how the interaction between network structure and synaptic plasticity realizes and sustains critical operation. Using compact spiking-network models that combine canonical topologies—small-world, scale-free, and modular—with self-tuning plasticity, we show how structural features and plasticity timescales expand or narrow the self-organized critical operating range in the vicinity of criticality.

Keywords: Brain criticality; Neuronal avalanches; Edge of chaos; Self-organized criticality (SOC).

Explaining Parasite Diversity and Host Diversity in Nature

Wei-chung Liu

Institute of Statistical Science, Academia Sinica, Taiwan

ABSTRACT

Parasites are ubiquitous in nature and they constitute a substantial amount of biomass in an ecosystem. Previous studies have shown parasitism is not a random process and there is a great heterogeneity in parasite diversity among host species. Furthermore, parasitism tends to occur in topologically important positions in an ecological network. To date, the mechanism explaining such phenomena remains largely elusive. To this end, in this study, we construct a simple simulation model to explain the observed patterns of parasitism in nature. Our model takes an ecological network, or a food web, as given, then it allows parasite species to randomly infect host species as initial seeds; then in a probabilistic manner, parasite species are transmitted via the predator-prey links throughout the food web. We then investigate whether the resulting patterns of parasitism, specifically the distributions of parasite diversity and host diversity, as well as the topological nature of those infected host species, are similar to their empirical counterparts. We also construct a random parasitism model for comparison purpose, and investigate the extent to which parasite transmission via predator-prey links can account for the observed patterns of parasitism in nature.

Keywords: host, parasite, diversity, transmission, model

Mapping Citation Tendencies among Broad Subject Groups: Closed Fields, Asymmetric Exchange, Unexpected Ties and Structural Disconnection

Wei-Chu Chiang[1], **Frederick Kin Hing Phoa[1*]**, Hiroka Hamada[2]

[1]*Institute of Statistical Science, Academia Sinica, Taiwan*

[2]*Institute of Statistical Mathematics, Japan*

ABSTRACT

This study maps the interaction structure among broad knowledge fields using citation data from the Web of Science. Building on a previously constructed Pointwise Mutual Information (PMI) matrix that captures citation probabilities among 3120 subject types, we propose an aggregation method to approximate citation tendencies at the broader subject-group level - applicable when raw citation frequencies are unavailable. The resulting matrix offers a condensed view focused on interactions among "pure" subject groups. Our analysis reveals a clear contrast between "closed-fields" (e.g., the humanities) and "open-fields" (e.g., biology-related fields), with notable exceptions such as chemistry and medicine. Citation tendencies are generally reciprocal, though several prominent asymmetric pairs indicate imbalanced flows of knowledge. We also identify a set of unexpectedly strong cross-group citation ties, often rooted in interdisciplinary research niches. In contrast, several subject group pairs exhibit unusually weak citation links, likely reflecting disciplinary specialization rather than unrealized opportunities for integration. We argue that the citation tendency matrix at the subject-group level serves as a valuable foundation for studying interdisciplinarity. It provides empirical insights into the structure of cross-field knowledge exchange and could be further applied to research metrics design or other practical applications.

Keywords: Web of Science; Pointwise Mutual Information (PMI); Interdisciplinarity; Citation Tendency; Knowledge Exchange

B-Spline Copula and Its Estimation

Xiaoling Dou[1*], Satoshi Kuriki[2], Gwo Dong Lin[3], Donald Richards[4]

[1]*Department of Natural Sciences, International Christian University*

[2]*The Institute of Statistical Mathematics*

[3]*Institute of Statistical Science, Academia Sinica*

[4]*Department of Statistics, Pennsylvania State University*

ABSTRACT

The B-spline copula is defined by a linear combination of elements of the normalized B-spline basis functions. The B-spline copula includes the Bernstein copulas as a special case. We examine the dependence properties of the B-spline copula and develop an EM algorithm to estimate the parameters of the copula. The EM algorithm is designed to maximize the penalized pseudo-likelihood function, wherein we use the smoothly clipped absolute deviation (SCAD) penalty function for the penalization term. We conduct simulation studies to demonstrate the stability of the proposed numerical procedure, show that penalization yields estimates with smaller mean-square errors when the true parameter matrix is sparse, and provide methods for determining tuning parameters and for model selection.

Keywords: B-spline basis functions; B-spline copula; EM algorithm; Model selection; SCAD penalty

Measuring Multivariate Regression Association via Spatial Sign

Jia-Han Shih[1,*], Yi-Hau Chen[2]

[1]*Department of Applied Mathematics, National Sun Yat-sen University*

[2]*Institute of Statistical Science, Academia Sinica*

ABSTRACT

In this talk, we propose a regression association measure aiming at the predictability of a multivariate outcome $\boldsymbol{Y}$ $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$from a multivariate covariate $\boldsymbol{X}$. We first generalize the conventional Kendall's tau to assess association between two random vectors. Then, we apply the generalized Kendall's tau to two independent replications from the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$, where $\boldsymbol{Y}$ and $\boldsymbol{Y'}$ share the same conditional distribution and conditionally independent given $\boldsymbol{X}$. The proposed measure can be expressed as the proportion of the variance of some function of $\boldsymbol{Y}$ $\boldsymbol{Y}$ that can be explained by $\boldsymbol{X}$, indicating that the measure has an interpretation in terms of predictability. Based on the proposed regression association measure, we further define a conditional regression association measure, which can be utilized to perform variable selection. Since our measure is based on two independent replications from the conditional distribution, a simple nonparametric estimation method based on nearest neighbor is available. Simulations are carried out to examine the performance of the proposed variable selection algorithm and real data examples are analyzed for illustration.

Keywords: Cosine similarity; Functional association; Kendall's tau; Variable selection

The Trivariate Wrapped Cauchy Copula

Shogo Kato[1], Christophe Ley[2], Sophia Loizidou[2], Kanti V. Mardia[3]

[1] *Institute of Statistical Mathematics*

[2] *Department of Mathematics, University of Luxembourg*

[3] *Department of Statistics, University of Leeds*

ABSTRACT

Toroidal data consist of observations comprising multiple angles, commonly found in environmental sciences such as wind directions and wave directions. In this talk, we propose a new distribution for trivariate toroidal data which we call a trivariate wrapped Cauchy copula. The proposed copula has the following advantages: (i) a simple form of density, (ii) an adjustable degree of dependence between every pair of variables, (iii) parameters with clear interpretation, (iv) well-known marginal and conditional distributions, (v) a straightforward data generating mechanism, (vi) unimodality, (vii) a closed-form expression for trigonometric moments, and (viii) a simple implementation procedure for obtaining maximum likelihood estimates. As is the case with general copula models, the proposed copula can be extended to have any specific marginal distributions and hence can be utilized for flexible modeling. Moreover, our construction allows for linear marginals, implying that our copula can also model cylindrical data, which consist of both angular and linear observations. As an application of the extended copula model, we consider a dataset of trivariate dihedral angles of amino acids in bioinformatics. Finally, we discuss how the proposed trivariate copula can be extended to multivariate copulas.

Keywords: Angular data; Directional statistics; Flexible modelling; Wrapped Cauchy distribution

A Bayesian Estimator of Sample Size

Dehua Bi[1], **Yuan Ji[1]**

[1]*Department of Public Health Sciences, The University of Chicago, IL*

ABSTRACT

We consider a Bayesian estimator of sample size (BESS) and an application to oncology dose optimization clinical trials. BESS is built upon three pillars, Sample size, Evidence from observed data, and Confidence in posterior inference. It uses a simple logic of "given the evidence from data, a specific sample size can achieve a degree of confidence in the posterior inference." The key distinction between BESS and standard sample size estimation (SSE) is that SSE, typically based on Frequentist inference, specifies the true parameters values in its calculation while BESS assumes possible outcome from the observed data. As a result, the calibration of the sample size is not based on type I or type II error rates, but on posterior probabilities. We demonstrate that BESS leads to a more interpretable statement for investigators, and can easily accommodates prior information as well as sample size re-estimation. We explore its performance in comparison to the standard SSE and demonstrate its usage through a case study of oncology optimization trial. BESS can be applied to general hypothesis tests. An R tool is available at https://ccte.uchicago.edu/BESS.

Keywords: Clinical trial; Confidence; Evidence; Hypothesis testing; Posterior inference; Priors; Type I error.

Sampling from the Random Linear Model via Stochastic Localization Up to the AMP Threshold

Han Cui, Zhiyuan Yu, **Jingbo Liu**

Department of Statistics, University of Illinois

ABSTRACT

Recently, Approximate Message Passing (AMP) has been integrated with stochastic localization (diffusion model) by providing a computationally efficient estimator of the posterior mean. Existing (rigorous) analysis typically proves the success of sampling for sufficiently small noise, but determining the exact threshold involves several challenges. In this work, we focus on sampling from the posterior in the linear inverse problem, with an i.i.d. random design matrix, and show that the threshold for sampling coincides with that of posterior mean estimation. We give a proof for the convergence in smoothed KL divergence whenever the noise variance is below the computation threshold for mean estimation introduced by (Barbier et al., 2020). We also show convergence in the Wasserstein distance under the same threshold assuming a dimension-free bound on the operator norm of the posterior covariance matrix, a condition strongly suggested by recent breakthroughs on operator norm bounds in similar replica symmetric systems. A key step in our analysis is to show that phase transition does not occur along the sampling and interpolation paths when the noise variance is below the computation threshold for mean estimation. We also discuss a new method for rigorously proving the consistency of an emerging Thouless-Anderson-Palmer (TAP) approach for mean estimation, which is believed to offer a more robust estimation than the AMP approach. (Based on arXiv:2407.10763 and arXiv:2506.20768)

Keywords: Posterior sampling; Bayesian estimation; Approximate message passing; M-estimation

Bayesian Smoothing and Feature Selection via Variational Automatic Relevance Determination

Zihe Liu, Diptarka Saha, **Feng Liang**

Department of Statistics, University of Illinois at Urbana-Champaign

ABSTRACT

This study introduces Variational Automatic Relevance Determination (VARD), a novel approach for fitting sparse additive regression models in high-dimensional settings. VARD stands out by independently assessing the smoothness of each feature while precisely determining whether its contribution to the response is zero, linear, or nonlinear. Additionally, we present an efficient coordinate descent algorithm for implementing VARD. Empirical evaluations on both simulated and real-world datasets demonstrate VARD's superior performance compared to alternative variable selection methods for additive models.

Keywords: Variational inference; Additive model; Smoothing; Feature selection

Prediction Interval Transfer Learning for Linear Regression Using an Empirical Bayes Approach

Anand Dixit[1], Weining Shen, **Min Zhang[3]**, Dabao Zhang[3]

[1]*Department of Statistics, Purdue University, West Lafayette, Indiana, USA*

[2]*Department of Statistics, University of California, Irvine, California, USA*

[3]*Department of Epidemiology and Biostatistics, University of California, Irvine, California, USA*

ABSTRACT

Current research in transfer learning focuses on improving the predictive performance for small datasets by leveraging information from larger, but potentially biased datasets. However, these methods do not provide prediction intervals, and as a result, one has to either rely solely on the small dataset or combine it with the possibly biased dataset to obtain prediction intervals using traditional linear regression methods. In this project, we propose a new approach, namely the Empirical Bayes approach for prediction interval transfer learning, to calculate prediction intervals within transfer learning for linear regression tasks. We have showed that the Gibbs sampler associated with our method is geometrically ergodic, which allows for the quantification of Monte Carlo uncertainty associated with its predicted values. In addition, the efficiency of our proposed approach is demonstrated through simulation studies and an application to a real-world dataset.

Keywords: Empirical Bayes; Prediction interval; Transfer learning

December 19 (Friday):

Parallel Sessions [10:20 – 12:00]:

19a1 - Statistica Sinica Special Invited Papers

19a2 - Frontiers in Statistical Inference: Dependence and Data Privacy

19a3 - Advances in Causality, Reinforcement Learning, and Business Analytics

19a4 - Advanced Methods for Novel Biomedical Data Types

19a5 - Recent Advancements in Network and Correlated Data Analysis

19a6 - Deep learning and Artificial Intelligence

19a7 - Recent Developments on Biostatistics and AI

19a8 - Recent Advances in Clinical Trials

19a9 - Recent Advance in Statistics and AI

19a10 - Modern Bayesian and Machine Learning Methods for Precision Medicine and Digital Health

Parallel Sessions [12:50 – 14:30]:

19b1 - What is Obscure about Random Objects?

19b2 - Data Visualization

19b3 - Recent Advances in Nonparametric Methods and Their Applications

19b4 - Extending Canonical Conformal Predictions to Meet the Practitioners' Needs

19b5 - Recent Research in Statistical Process Control, Part II

19b6 - Recent Development in AI

19b7 - Recent Development of Statistical Methods in Case-Cohort Study Design and Dependent Sampling

19b8 - Recent Research Developments in Neuroimaging Data Analysis

19b9 - Causal Inference: Episode I

19b10 - Recent Developments in Survival Analysis and Clinical Trial Methodology

The Method of Limits and Its Application to The Analysis of Count Data in Genome-Wide Association Studies

Jiming Jiang[1,*], Leqi Xu[2], Yiliang Zhang[2] and Hongyu Zhao[2]

[1]*University of California, Davis and* [2]*Yale University*

ABSTRACT

We propose a new method of statistical inference, called the method of limits (MoL), which may be viewed as an extension of the method of moments. This method is motivated by the need to analyze count data for genome wide association studies (GWAS), where the existing methods are hindered in statistical inference due to computational challenges. We establish consistency and asymptotic normality of the MoL estimator of heritability from GWAS data, which is seen as an advantage over the existing PQLseq method. Furthermore, we derived a consistent estimator of the proportion of causal SNPs. MoL also showed an advantage of both statistical and computational efficiency measured by average statistical efficiency (ASE) in our simulation studies compared to PQLseq. We also illustrate the usefulness of MoL through its application to the UK Biobank data to infer the heritability of weekly champagne consumption and week red wine consumption using the count data.

Keywords: Asymptotic properties; Big GWAS data; Proportion of causal SNPs; Relative average statistical efficiency

Weighted Conditional Network Testing for Multiple High-Dimensional Correlated Data Sets

Takwon Kim[1], **Inyoung Kim[2,*]** , Ki-Ahm Lee[3]

Department of Statistics, Virginia Tech

ABSTRACT

Gaussian graphical models (GGMs) have been investigated to infer dependence (or network) structure among high-dimensional data by estimating a precision matrix. However, while many estimation methods for GGM have been developed, methods for testing the equality of two precision matrices are still limited. Because testing the equality of the precision matrix depends on other given precision matrices, we develop a weighted conditional network testing for considering other given precision matrices information and also provides theoretical properties. None of the existing methods can be applied to test conditional differences when other networks are conditionally given and different. We demonstrate the advantage of our approach using a simulation study and genetic pathway analysis.

Keywords: Conditional Difference; Gaussian Graphical Model; Precision Matrix

Statistical Inference for Differentially Private Stochastic Gradient Descent

Zhanrui Cai[1]

Faculty of Business and Economics, University of Hong Kong

ABSTRACT

Privacy preservation in machine learning, particularly through Differentially Private Stochastic Gradient Descent (DP-SGD), is critical for sensitive data analysis. However, existing statistical inference methods for SGD predominantly focus on cyclic subsampling, while DP-SGD requires randomized subsampling. This paper first bridges this gap by establishing the asymptotic properties of SGD under the randomized rule and extending these results to DP-SGD. For the output of DP-SGD, we show that the asymptotic variance decomposes into statistical, sampling, and privacy-induced components. Two methods are proposed for constructing valid confidence intervals: the plug-in method and the random scaling method. We also perform extensive numerical analysis, which shows that the proposed confidence intervals achieve nominal coverage rates while maintaining privacy.

Keywords: Statistical inference; Stochastic gradient descent; Differential privacy

Greedy Model Selection under Sparsity and Covariate Shift

Ching-Kang Ing

Institute of Statistics and Data Science, National Tsing Hua University

ABSTRACT

Sparsity assumptions are central to high-dimensional model selection, yet the true sparsity level is typically unknown in practice. We investigate the convergence of the Chebyshev Greedy Algorithm (CGA) under weak sparsity conditions and propose a data-driven stopping rule, based on a high-dimensional information criterion (HDIC), to adaptively determine the number of iterations. After briefly noting that the CGA+HDIC framework attains the optimal convergence rate without prior knowledge of sparsity, we turn to the more challenging setting where covariates are subject to distributional change. To address this, we extend CGA to an importance-weighted version (IWCGA) by incorporating importance weights into the algorithm. In parallel, we develop an importance-weighted variant of HDIC, termed HDIWIC, to improve model selection under distribution shift. We establish convergence guarantees for the joint IWCGA-HDIWIC procedure and demonstrate its effectiveness through simulations and real-data applications.

Keywords: Chebyshev Greedy Algorithm; Covariate shift; High-dimensional model selection; Importance weighting

Tying Maximum Likelihood Estimation for Dependent Data

Masamune Iwasawa[1], **Qingfeng Liu[2]**, Ziyan Zhao[3]

[1]*Doshisha University, Faculty of Economics*

[2]*Department of Industrial and Systems Engineering, Hosei University*

[3]*Management School, Lancaster University*

ABSTRACT

This study proposes a tying maximum likelihood estimation (TMLE) method to improve the estimation performance of parametric models for dependent data where some time series have long sample periods, while the others are significantly shorter. The TMLE achieves this by flexibly tying some parameters of the long time series to those of the short ones, facilitating the transfer of valuable information to improve the parameter estimation accuracy for the short series. We derive the asymptotic properties of the TMLE and its finite-sample risk bound under a tuning parameter that determines the strength of the tying. The theoretical analysis shows that the TMLE not only substantially outperforms the standard MLE under the correct tying but also can maintain this strength under a local misspecification setting. We propose a feasible bootstrapping procedure for selecting the tuning parameter to reduce the finite-sample risk, with a supporting finite-sample theory that can guide effective implementation of the procedure. Extensive simulations and empirical applications demonstrate that the TMLE exhibits superior performance compared to alternative methods.

Keywords: Tying; MLE; Finite-sample theory; Local misspecification

LLM-Powered Prediction Inference with Online Text Time Series

Yingying Fan[1], **Jinchi Lv[1]**, Ao Sun[1] and Yurou Wang[2]

[1]*University of Southern California*

[2]*Xiamen University*

ABSTRACT

Time series prediction inference is an important yet challenging task in economics and business, where existing approaches often rely on low-frequency, survey-based data. With the recent advances of large language models (LLMs), there is growing potential to leverage high-frequency online text data for improved time series prediction, an area still largely unexplored. This paper proposes LLM-TS, an LLM-based approach for time series prediction inference incorporating online text data. The LLM-TS is based on a joint time series framework that combines survey-based low-frequency data with LLM-generated high-frequency surrogates. The framework relies only on an error correlation assumption, combining a text-embedding-augmented ARX model for the observed gold-standard measurements with a VARX model for the LLM-generated surrogates. LLM-TS employs LLMs such as ChatGPT and the trained BERT models to construct LLM surrogates. Online text embeddings are extracted via LDA and BERT. We establish the asymptotic properties of the method and provide two forms of constructed prediction intervals. To demonstrate the practical power of LLM-TS, we apply it to a critical real-world example: inflation forecast. We collect a large set of high-frequency online texts from a widely used Chinese social media platform and employ LLMs to construct inflation labels for posts that are related to inflation. The finite-sample performance and practical advantages of LLM-TS are illustrated through simulations and this noisy real data example, highlighting its potential to improve time series prediction in economic applications. This is a joint work with Yingying Fan, Ao Sun and Yurou Wang.

Keywords: large language models, CPI prediction, Online texts, Asymptotic distributions, Time series

Learning Robust Decision Rules for Censored and Confounded Data

Yifan Cui

Zhejiang University

ABSTRACT

In this talk, we propose two robust criteria for learning optimal treatment rules with censored survival outcomes. The first one aims to identify a treatment rule that maximizes the restricted mean survival time, where the restriction is specified by a given quantile such as the median; the second one focuses on maximizing buffered survival probabilities, with the threshold adaptively adjusted to account for the restricted mean survival time. Moreover, we develop robust treatment rules that enable reliable policy recommendations when unmeasured confounding is present, using the proximal causal inference framework. Simulation studies and real-world applications demonstrate the superior performance of the proposed methods.

Keywords: Causal Inference, Decision-making, Survival analysis

Causal Inference for All: Marginal Causal Effects for Outcomes Truncated by Death

Ruixuan Zhao[1], Mats Stensrud[2], **Linbo Wang[1,3*]**

[1]*Department of Computer and Mathematical Sciences, University of Toronto*

[2]*Institute of Mathematics, EPFL*

[3]*Department of Statistical Sciences, University of Toronto*

ABSTRACT

In longitudinal studies where outcomes may be truncated by death, standard causal estimands often fail to capture meaningful treatment effects, particularly when survival is affected by treatment. Traditional survivor average causal effects (SACEs), which condition on post-treatment survival, are challenging to interpret and identify without strong assumptions, and their direct extension to longitudinal settings poses additional difficulties. We propose a flexible class of marginal causal estimands that aggregate potential outcomes over time among individuals who would survive under both treatment and control. These estimands support a range of clinically relevant summaries, such as cumulative or last-observed outcomes, and can be tailored using weighting schemes to align with different decision-making goals. We illustrate these ideas through a reanalysis of a prostate cancer clinical trial, highlighting how different estimands may lead to different treatment conclusions.

Keywords: Local average effects, Longitudinal data analysis, Missing data, Selection bias

Quantum Speedups for Multiproposal MCMC

Andrew Holbrook
UCLA

ABSTRACT

:

Multiproposal MCMC algorithms choose from multiple proposals to generate their next chain step in order to sample from challenging target distributions more efficiently. However, on classical machines, these algorithms require O(P) target evaluations for each Markov chain step when choosing from P proposals. Recent work demonstrates the possibility of quadratic quantum speedups for one such multiproposal MCMC algorithm. After generating P proposals, this quantum parallel MCMC (QPMCMC) algorithm requires only O(\sqrt{P}) target evaluations at each step, outperforming its classical counterpart. Here, I present a new, faster quantum multiproposal MCMC strategy, QPMCMC2. With a specially designed proposal distribution, QPMCMC2 requires only O(1) target evaluations and O(log P) qubits when computing over a large number of proposals P. Unlike its slower predecessor, the QPMCMC2 Markov kernel (1) maintains detailed balance exactly and (2) is fully explicit for a large class of graphical models. I demonstrate this flexibility by applying QPMCMC2 to novel Ising-type models built on bacterial evolutionary networks and obtain significant speedups for Bayesian ancestral trait reconstruction for 248 observed salmonella bacteria.

Keywords: Multiproposal MCMC; Quantum Computing; Ising models

Indirect Statistical Inference with Guaranteed Necessity and Sufficiency

Zhengjun Zhang[1,2,3], Xinyang Hu[4], Chuyang Lu[5], and Tianying Liu[1]

[1]*School of Economics and Management, University of Chinese Academy of Sciences*

[2]*AMSS Center for Forecasting Science, Chinese Academy of Sciences*

[3]*Department of Statistics, University of Wisconsin, Madison*

[4]*Department of Statistics, Yale University*

[5]*Academy of Mathematics and System Sciences, Chinese Academy of Sciences*

ABSTRACT

This paper develops a new framework for indirect statistical inference with guaranteed necessity and sufficiency, applicable to continuous random variables. We prove that when comparing exponentially transformed order statistics from an assumed distribution with those from simulated unit exponential samples, the ranked quotients exhibit distinct asymptotics: the left segment converges to a non-degenerate distribution, while the middle and right segments degenerate to one. This yields a necessary and sufficient condition in probability for two sequences of continuous random variables to follow the same distribution. Building on this, we propose an optimization criterion based on relative errors between ordered samples. The criterion achieves its minimum if and only if the assumed and true distributions coincide, providing a second necessary and sufficient condition in optimization. These dual NS properties, rare in the literature, establish a fundamentally stronger inference framework than existing methods. Unlike classical approaches based on absolute errors (e.g., Kolmogorov–Smirnov), NSE exploits relative errors to ensure faster convergence, requires only mild approximability of the cumulative distribution function, and provides both point and interval estimates. Simulations and real-data applications confirm NSE's superior performance in preserving distributional assumptions where traditional methods fail.

Keywords: combinatorial mathematics, indirect inference, relative errors, simulated order statistics.

BrainGeneBot: A GPT-Engineered, User-Driven Genetic Data Exploration with Polygenic Risk Scores Ranking in Alzheimer's Disease

Zhongming Zhao, PhD

McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston

ABSTRACT

Polygenic risk scores (PRS) are widely used to assess genetic susceptibility in Alzheimer's Disease (AD) research. However, the rapid expansion of PRS studies has led to dataset-specific biases leading to inconsistent variant prioritization and limit generalizability and reproducibility. To address these challenges, we propose a transductive learning framework that integrates multiple PRS datasets for more robust risk variant prioritization, incorporating Genome-Wide Association Study (GWAS) priority scores as biologically informed priors. Additionally, we introduce BrainGeneBot, an AI-driven tool leveraging Generative Pre-trained Transformers (GPT) with Retrieval-Augmented Generation (RAG) technology to streamline genomic analyses in AD, including the STRING for protein interaction analysis, Enrichr for gene set enrichment, ClinVar for genetic variant interpretation, and Biopython for conducting literature searches. We apply our approach to publicly available AD datasets from the PGS Catalog and conduct further analyses to validate its efficacy. In parallel, we perform conventional unsupervised rank aggregation as a baseline. The transductive learning approach not only verifies high-risk variants identified by traditional methods but also reveals unique insights that better correlate with GWAS signals. Our framework streamlines data retrieval and interpretation, effectively prioritizing genetic variants in multiple PRS studies. In summary, the implementation of BrainGeneBot is set to transform genomic research for brain diseases by improving data accessibility, accelerating discovery processes, and refining the precision of genetic insights.

Keywords: GPT-powered informatics, polygenic score, rank aggregation, transductive learning

High-Dimensional Markov-Switching Ordinary Differential Processes

Katherine Tsai[1,3*], Mladen Kolar[2], Sanmi Koyejo[3]

[1]*Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign*

[2]*Marshall School of Business, University of Southern California*

[3]*Department of Computer Science, Stanford University*

ABSTRACT

We investigate the parameter recovery of Markov-switching ordinary differential processes from discrete observations, where the differential equations are nonlinear additive models. This framework has been widely applied in biological systems, control systems, and other domains; however, limited research has been conducted on reconstructing the generating processes from observations. In contrast, many physical systems, such as human brains, cannot be directly experimented upon and rely on observations to infer the underlying systems. To address this gap, this manuscript presents a comprehensive study of the model, encompassing algorithm design, optimization guarantees, and quantification of statistical errors. Specifically, we develop a two-stage algorithm that first recovers the continuous sample path from discrete samples and then estimates the parameters of the processes. We provide novel theoretical insights into the statistical error and linear convergence guarantee when the processes are beta-mixing. Our analysis is based on the truncation of the latent posterior processes and demonstrates that the truncated processes approximate the true processes under mixing conditions. We apply this model to investigate the differences in resting-state brain networks between the ADHD group and normal controls, revealing differences in the transition rate matrices of the two groups.

Keywords: Ordinary Differential Processes; Regime Switchings; Latent Models

Controlling False Discover Rate for High Dimensional Mediator Selection in Non-linear Models

Runqiu Wang[1], **Ran Dai[1]**, Jieqiong Wang[2], Kah Meng Soh[1], Ziyang Xu[2], Mohamed Azzam[2], Hongying Dai[1], Cheng Zheng[1]

[1]Department of Biostatistics, University of Nebraska Medical Center, 984375 Nebraska Medical Center, 68198, Nebraska, U.S.A.

[2]Department of Neurological Sciences, University of Nebraska Medical Center, 984375 Nebraska Medical Center, 68198, Nebraska, U.S.A.

ABSTRACT

There is a challenge in selecting high-dimensional mediators when the mediators have complex correlation structures and interactions. In this work, we frame the high-dimensional mediator selection problem into a series of hypothesis tests with composite nulls, and develop a method to control the false discovery rate (FDR) which has mild assumptions on the mediation model. We show the theoretical guarantee that the proposed method and algorithm achieve FDR control. We present extensive simulation results to demonstrate the power and finite sample performance compared with existing methods. Lastly, we demonstrate the method for analyzing the Alzheimer's Disease Neuroimaging Initiative (ADNI) data, in which the proposed method selects the volume of the hippocampus and amygdala, as well as some other important MRI-derived measures as mediators for the relationship between gender and dementia progression.

Keywords: False Discovery Rate; high-dimensional mediators; imaging data; knockoff

When Few Labeled Target Data Suffice: A Theory of Semi-Supervised Domain Adaptation via Fine-Tuning from Multiple Adaptive Starts

Wooseok Ha[1], Yuansi Chen[2]

[1]*Department of Mathematical Sciences, KAIST*

[2]*Department of Mathematics, ETH Zu¨rich*

ABSTRACT

Semi-supervised domain adaptation (SSDA) aims to achieve high predictive performance in the target domain with limited labeled target data by exploiting abundant source and unlabeled target data. Despite its significance in numerous applications, theory on the effectiveness of SSDA remains largely unexplored, particularly in scenarios involving various types of source-target distributional shifts. In this talk, I will present a theoretical framework based on structural causal models (SCMs) which allows us to analyze and quantify the performance of SSDA methods when labeled target data is limited. Within this framework, I introduce three SSDA methods, each having a fine-tuning strategy tailored to a distinct assumption about the source and target relationship. Under each assumption, I demonstrate how extending an unsupervised domain adaptation (UDA) method to SSDA can achieve minimax-optimal target performance with limited target labels. Finally, when the relationship between source and target data is only vaguely known—a common practical concern—I will describe the Multi Adaptive-Start FineTuning (MASFT) algorithm, which fine-tunes UDA models from multiple starting points and selects the best-performing one based on a small hold-out target validation dataset. Combined with model selection guarantees, MASFT achieves near-optimal target predictive performance across a broad range of types of distributional shifts while significantly reducing the need for labeled target data.

Keywords: Semi-supervised domain adaptation; distribution shifts; fine-tuning; invariance

Weining Wang

Covariance-Based Clustering and Biclustering via the Heterogeneous Block Covariance Model and Variants

Yunpeng Zhao, Xiang Li, Ning Hao, Qing Pan, and Mayson Zhang

Department of Statistics, Colorado State University

ABSTRACT

Clustering methods are traditionally designed to group data points in a (possibly high-dimensional) Euclidean space. We study a different but equally important task: clustering at the feature level, which arises naturally in gene expression analysis yet has received limited attention in statistics. In bioinformatics, it is often desirable to identify sets of highly correlated genes whose expression levels are jointly regulated and act synergistically to perform the same biological function. To address this problem, we introduce the heterogeneous block covariance model (HBCM), which characterizes community structure directly within the covariance matrix while accounting for heterogeneity in how features connect within communities. We develop a novel variational expectation–maximization algorithm for efficient estimation of group memberships. Theoretical analysis establishes the consistency of membership recovery, and simulation studies demonstrate the superior performance of HBCM compared to existing methods. We further extend HBCM to a joint clustering and biclustering framework that simultaneously partitions both features and samples. This generalization introduces new computational and theoretical challenges, for which we propose principled solutions. Applications to real gene expression data highlight the model's ability to uncover biologically meaningful clusters and biclusters.

Keywords: Variational EM algorithm; Covariance matrix; Biclustering; Gene expression data

Adaptive Block-Based Change-Point Detection for Sparse Spatially Clustered Data with Applications in Remote Sensing Imaging

Alan Moore, **Lynna Chu**, Zhengyuan Zhu

Department of Statistics, Iowa State University

ABSTRACT

We present a non-parametric change-point detection approach to detect potentially sparse changes in a time series of high-dimensional observations or non-Euclidean data objects. We target a change in distribution that occurs in a small, unknown subset of dimensions, where these dimensions may be correlated. Our work is motivated by a remote sensing application, where changes occur in small, spatially clustered regions over time. An adaptive block-based change-point detection framework is proposed that accounts for spatial dependencies across dimensions and leverages these dependencies to boost detection power and improve estimation accuracy. Through simulation studies, we demonstrate that our approach has superior performance in detecting sparse changes in datasets with spatial or local group structures. An application of the proposed method to detect activity, such as new construction, in remote sensing imagery of the Natanz Nuclear facility in Iran is presented to demonstrate the method's efficacy.

Keywords: Change-point; Non-parametric; Graph-based tests; Spatial dependence; Satellite images

Graph Release with Assured Node Differential Privacy

Suqing Liu, Xuan Bi, **Tianxi Li**[1]

University of Chicago and University of Minnesota, Twin Cities

ABSTRACT

Differential privacy is a well-established framework for safeguarding sensitive information in data. While extensively applied across various domains, its application to network data --- particularly at the node level --- remains underexplored. Existing methods for node-level privacy either focus exclusively on query-based approaches, which restrict output to pre-specified network statistics, or fail to preserve key structural properties of the network. In this work, we propose GRAND (Graph Release with Assured Node Differential privacy), which is, to the best of our knowledge, the first network release mechanism that releases entire networks while ensuring node-level differential privacy and preserving structural properties. Under a broad class of latent space models, we show that the released network asymptotically follows the same distribution as the original network. The effectiveness of the approach is evaluated through extensive experiments on both synthetic and real-world datasets.

Keywords: Data Privacy; Privacy-Preserving Networks; Data Sharing; Network Release

Advancing Responsible Statistical and AI/ML Methods for Harnessing the Power of Electronic Health Records

Qi Long

Division of Biostatistics, University of Pennsylvania

ABSTRACT

Rich electronic health records (EHR) data offer remarkable opportunities in advancing precision medicine (Orcutt et al., 2025), they also present daunting analytical challenges. Multi-modal data in EHR that are recorded at irregular time intervals with varying frequencies include structured data such as labs and vitals, codified data such as diagnosis and procedure codes, and unstructured data such as clinical notes and pathology reports. They are typically incomplete and fraught with other errors and biases. What's more, data gaps and errors in EHRs are often unequally distributed across patient groups: People with less access to care, often people with lower socioeconomic status, tend to have more incomplete data in EHRs. Such data issues, if not adequately addressed, would lead to biased results and exacerbate health inequities (Getzen et al. 2023). In this talk, I will share my research group's recent works on developing responsible statistical and AI/ML methods including large language models (LLMs) for addressing these challenges (Zhang et al., 2024; Consoli et al., 2024). Since LLMs are themselves plagued by various biases, I will also discuss our ongoing research on developing rigorous statistical and machine learning methods for mitigating pitfalls and risks of LLMs.

Keywords: AI/ML; Electronic Health Records; Large Language Models; Precision Medicine.

Deep Survival Analysis for Competing Risk Modeling with Functional Covariates and Missing Data Imputation

Xiaofeng Wang, PhD[1], Penglei Guo, PhD,[1] Yan Zou, MS,[1] Abhijit Duggal, MD[2], Shuaiqi Huang, PhD[1], Faming Liang, PhD[3]

[1]*Department of Quantitative Health Science, Cleveland Clinic, Cleveland OH, USA,*

[2]*Department of Pulmonary and Critical Care Medicine, Cleveland Clinic, Cleveland OH, USA,*

[3]*Department of Statistics, Purdue University, Lafayette IN, USA*

ABSTRACT

We introduce the Functional Competing Risk Net (FCRN), a unified deep-learning framework for discrete-time survival analysis under competing risks, seamlessly integrating functional covariates and handling missing data within an end-to-end model. By combining a micro-network Basis Layer for functional data representation with a gradient-based imputation module, FCRN simultaneously learns to impute missing values and predict event-specific hazards. Evaluated on multiple simulated datasets and a real-world ICU case study, FCRN demonstrates substantial improvements in prediction accuracy over random survival forests and traditional competing risks models. This approach advances prognostic modeling in critical care by effectively capturing dynamic risk factors and static predictors while accommodating irregular and incomplete data.

Keywords: Deep Learning; Competing Risks; Functional Data; Gradient-based Imputation

Mini-Batch Estimation for Deep Cox Models: Statistical Foundations and Practical Guidance

Ying Ding[1, *], Lang Zeng[1], Weijing Tang[2], Zhao Ren[3]

[1]*Department of Biostatistics and Health Data Science, University of Pittsburgh*

[2]*Department of Statistics and Data Science, Carnegie Mellon University*

[3]*Department of Statistics, University of Pittsburgh*

ABSTRACT

The stochastic gradient descent (SGD) algorithm has been widely used to optimize deep Cox neural network (Cox-NN) by updating model parameters using mini-batches of data. We show that SGD aims to optimize the average of mini-batch partial-likelihood, which is different from the standard partial-likelihood. This distinction requires developing new statistical properties for the global optimizer, namely, the mini-batch maximum partial-likelihood estimator (mb-MPLE). We establish that mb-MPLE for Cox-NN is consistent and achieves the optimal minimax convergence rate up to a polylogarithmic factor. For Cox regression with linear covariate effects, we further show that mb-MPLE is √n-consistent and asymptotically normal with asymptotic variance approaching the information lower bound as batch size increases, which is confirmed by simulation studies. Additionally, we offer practical guidance on using SGD, supported by theoretical analysis and numerical evidence. For Cox-NN, we demonstrate that the ratio of the learning rate to the batch size is critical in SGD dynamics, offering insight into hyperparameter tuning. For Cox regression, we characterize the iterative convergence of SGD, ensuring that the global optimizer, mb-MPLE, can be approximated with sufficiently many iterations. Finally, we demonstrate the effectiveness of mb-MPLE in a large-scale real-world application where the standard MPLE is intractable.

Keywords: Linear scaling rule; minimax rate of convergence; Stochastic gradient descent; Survival analysis

A Deep Learning Feature Importance Test for Integrating Informative High-dimensional Biomarkers

Baiming Zou[1,2], James G. Xenakis[3], Meisheng Xiao[1], Apoena Ribeiro[4], Kimon Divaris[4], Di Wu[1,4], Fei Zou[1,5]

[1]*Department of Biostatistics, Giling School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA;*

[2]*School of Nursing, University of North Carolina, Chapel Hill, NC, USA;*

[3]*Department of Statistics, Harvard University, Cambridge, MA, USA;*

[4]*School of Dentistry, University of North Carolina, Chapel Hill, NC, USA;*

[5]*Department of Genetics, School of Medicine, University of North Carolina, Chapel Hill, NC, USA.*

ABSTRACT

Many human diseases result from a complex interplay of behavioral, clinical, and molecular factors. Integrating low-dimensional behavioral and clinical features with high-dimensional molecular profiles can significantly improve disease outcome prediction and diagnosis. However, while some biomarkers are crucial, many lack informative value. To enhance prediction accuracy and understand disease mechanisms, it is essential to integrate relevant features and identify key biomarkers, separating meaningful data from noise and modeling complex associations. To address these challenges, we introduce the high-dimensional feature importance test (HdFIT) framework for machine learning models. HdFIT includes a feature screening step for dimension reduction and leverages machine learning to model complex associations between biomarkers and disease outcomes. It robustly evaluates each feature's impact. Extensive Monte Carlo experiments and a real microbiome study demonstrate HdFIT's efficacy, especially when integrated with advanced models like deep neural networks (DNN), termed HdFIT-DNN. Our framework shows significant improvements in identifying crucial features and enhancing prediction accuracy.

Keywords: Complex association, Dimension reduction, Interpretable and scalable predictive modeling, Non-parametric feature selection, Stable deep neural network

Presenter's email address: bzou@email.unc.edu

Latent Noise Injection for Private and Statistically Aligned Synthetic Data Generation

Rex Shen[1], **Lu Tian[2]**

[1]*Google Inc.z*

[2] *Stanford University*

ABSTRACT

Synthetic data generation has become essential for scalable, privacy-preserving statistical analysis. While standard generative-model–based approaches— such as those using Normalizing Flows— are widely adopted, they often exhibit slow convergence in high-dimensional settings, frequently failing to achieve the canonical root n rate when approximating the true data distribution. To address these limitations, we propose a *Latent Noise Injection* method built on Masked Autoregressive Flows. Rather than sampling directly from the trained model, our approach perturbs each observed data point in the latent space and maps it back to the data domain. This construction preserves a one-to-one correspondence between observed and synthetic samples, enabling synthetic outputs that more faithfully reflect the underlying distribution—particularly in challenging high-dimensional regimes where traditional sampling deteriorates. Our procedure satisfies local (ε,δ)-differential privacy and introduces a single perturbation parameter governing the privacy–utility trade-off. Although estimators based on a single synthetic dataset may converge slowly, we show both theoretically and empirically that aggregating results across multiple studies in a meta-analytic framework restores classical efficiency and yields consistent, reliable inference. With an appropriately calibrated perturbation parameter, Latent Noise Injection achieves strong statistical fidelity to the original data while providing robustness against membership-inference attacks. These results position our method as a compelling alternative to conventional flow-based sampling for synthetic data sharing in decentralized and privacy-sensitive domains, such as biomedical research.

Keywords: Synthetic data, privacy protection, normalizing flows.

MR2G: A Novel Framework for Causal Network Inference Using GWAS Summary Data

Zhaotong Lin[1], Wei Pan[2], **Haoran Xue[3]**

[1]*Department of Statistics, Florida State University*

[2]*Division of Biostatistics & Health Data Science, University of Minnesota*

[3]*Department of Biostatistics, City University of Hong Kong*

ABSTRACT

Inferring a causal network among multiple traits is essential for unravelling complex biological relationships and informing interventions. Mendelian randomization (MR) has emerged as a powerful tool for causal inference, utilizing genetic variants as instrumental variables (IVs) to estimate causal effects. However, when the directions of causal relationships among traits are unknown, reconstructing the underlying causal network becomes challenging. In particular, the presence of cycles or feedback loops, which are common in biological systems, poses additional challenges for causal network inference, and remains largely under-studied with standard MR approaches and existing IV-based network inference methods. To address these issues, we introduce MR2G, a new statistical framework that enables robust inference of causal networks, including those with cycles, directly from GWAS summary statistics. MR2G is built on a formally defined recursive causal graph model that rigorously links direct causal effects to MR estimands. It recovers a biologically interpretable causal network from pairwise MR effect estimates, while incorporating a network-informed IV screening strategy to reduce pleiotropic bias and improve robustness. Through realistic simulations, MR2G demonstrates superior accuracy and robustness in recovering complex causal structures, including those involving feedback loops. We apply MR2G to GWAS summary statistics for six complex diseases and nine cardiometabolic risk factors. MR2G not only recovers well-established causal pathways but also uncovers multiple feedback relationships, highlighting its utility in disentangling complex and biologically plausible causal networks from large-scale genetic data.

Keywords: Mendelian randomization; cyclic causal network; instrumental variable.

Modeling Non-Uniform Hypergraphs Using Determinantal Point Processes

Ji Zhu

University of Michigan

ABSTRACT

Most statistical models for networks focus on pairwise interactions between nodes. However, many real-world networks involve higher-order interactions among multiple nodes, such as co-authors collaborating on a paper. Hypergraphs provide a natural representation for these networks, with each hyperedge representing a set of nodes. The majority of existing hypergraph models assume uniform hyperedges (i.e., edges of the same size) or rely on diversity among nodes. In this work, we propose a new hypergraph model based on non-symmetric determinantal point processes. The proposed model naturally accommodates non-uniform hyperedges, has tractable probability mass functions, and accounts for both node similarity and diversity in hyperedges. For model estimation, we maximize the likelihood function under constraints using a computationally efficient projected adaptive gradient descent algorithm. We establish the consistency and asymptotic normality of the estimator. Simulation studies confirm the efficacy of the proposed model, and its utility is further demonstrated through edge predictions on several real-world datasets.

Keywords: Hypergraph; Determinantal Point Process; Network.

Patient-Centric Pragmatic Trials: Opening the DOOR to Benefit: Risk-Based Evaluation

Toshimitsu Hamasaki[1,*], Scott R. Evans[1]

[1]*The Biostatistics Center and Department of Biostatistics and Bioinformatics, Milken Institute School of Public Heath, The George Washington University*

ABSTRACT

Randomized clinical trials are considered the gold standard for evaluating the benefits and harms of therapeutic interventions. However, they often fall short in providing the evidence needed to meaningfully inform clinical decision-making. A major reason is the failure to identify and prioritize the most clinically relevant questions to be addressed through the trial. Traditional approaches frequently overlook these questions, leading to study designs, monitoring plans, analyses, and reporting that are misaligned with the actual needs of patient care. In particular, the conventional analytical practice of evaluating outcomes separately presents several limitations: it does not account for the associations or cumulative nature of multiple outcomes in individual patients; it complicates interpretation due to competing risks; it overlooks meaningful gradations in patient-centric outcomes; and it often separates efficacy and safety analyses across different populations, making benefit: risk assessment and generalizability unclear.

To address these challenges, the Desirability of Outcome Ranking (DOOR) paradigm has been developed and implemented in clinical trials and other medical research settings. The DOOR offers a patient-centric framework for trial design, data monitoring, analysis, interpretation, and reporting, with a focus on a comprehensive evaluation of the benefit–risk profile of therapeutic interventions. The paradigm shifts the emphasis toward the overall clinical desirability of outcomes for individual patients.

This presentation: (1) provides a comprehensive framework for the DOOR paradigm, covering statistical methods for the design and analysis of clinical trials, guiding principles, and a recommended statistical analysis plan (SAP), and (2) highlights key challenges in the design and analysis of clinical trials, illustrating how the DOOR approach addresses the challenges and offers practical, patient-centric solutions.

Keywords: Benefit: risk; Clinical Trials; Desirability of Outcome Ranking (DOOR); Partial credit; Patient-centric; Pragmatic trials; Rank-based Analysis

Adaptive Population Selection Designs for Clinical Trials with Multiple Endpoints

Koko Asakura[1], Toshimitsu Hamasaki[2], Frank Bretz[3]

[1]*National Cerebral and Cardiovascular Center, Osaka, Japan*

[2]*The George Washington University Biostatistics Center, Bethesda, USA*

[3]*Novartis Pharma AG, Basel, Switzerland*

ABSTRACT

In some clinical trial settings, there may be uncertainty about which patient populations are most likely to benefit from a new treatment. For example, a treatment may be hypothesized to be effective in a prespecified targeted subgroup—defined by demographic characteristics, genetic markers, or pathophysiological features relevant to the treatment's mechanism of action—while its benefit in the non-targeted (complementary) subgroup remains unclear.

This presentation investigates confirmatory clinical trials that allow adaptive selection of patient populations with prespecified subgroups based on unblinded interim results, and to assess the treatment effect in the selected subgroups for multiple endpoints. Specifically, enrollment may continue in the overall population or be restricted to the targeted subgroup. At the end of the trial, the treatment effect may then be evaluated in the overall population, the targeted subgroup, or both. In such settings, data collected before and after the interim analysis must be appropriately integrated to support valid inference on treatment effects, accounting for the multiple endpoints and potential adaptations.

This setting poses several methodological challenges: (1) the complexity of ensuring consistent patient population selection across multiple endpoints at an interim analysis; and (2) the need to consider multiple sources of multiplicity due to the involvement of multiple endpoints, subgroups, and analyses. We address both co-primary endpoints and hierarchically ordered endpoints (e.g., primary and secondary). Our approach builds on the closure principle, and multiple hypotheses are assessed using combination functions to test intersection hypotheses within the closed testing procedure, alongside specific test procedures and interim decision rules for co-primary and hierarchically ordered endpoints. We evaluate the performance of the proposed methodology, including power and sample size under Type I error control, through simulation studies. We illustrate these approaches with an example.

Keywords: Biomarker; Subgroup; Type I error; Interim analysis; Multiple Endpoint

Sample Size Assessment for Survival Trial Designs with Covariate-Adaptive Randomization

Pei-Fang Su[1*], Chieh-Chi Wu[1]

[1]*Department of Statistics, National Cheng Kung University, Tainan, Taiwan*

ABSTRACT

Estimating the required sample size is an essential task for a clinical trial. It ensures statistical power and makes it possible to draw meaningful conclusions from the study results. Moreover, to minimize treatment imbalances within each covariate subgroup, covariate-adaptive randomization is a popular method. The aim of this study is to investigate the required sample size for covariate-adaptive randomization based on survival outcomes. We evaluate the testing performance using the calculated sample size under simple randomization, stratified permuted block randomization, and covariate-adaptive biased coin randomization. In order to provide preliminary insights into the trial's progress and potential efficacy, an interim analysis is commonly conducted.

The second aim of the study is to provide a strategy for interim analysis in covariate-adaptive randomization trials, which involves stopping the trial early based on accumulating data if one treatment arm proves to be significantly more effective or harmful than the others. We thus re-estimate the required sample size, propose a valid hypothesis testing method for interim analysis, and study the underlying theoretical properties of the testing statistics incorporating covariate-adaptive randomization. The performance of the proposed formula and interim strategy is evaluated through comprehensive simulations, including a sensitivity analysis. We also provide the R code to benefit the readers. Finally, an example is treated as a pilot studie, and we show that the proposed strategies are valid under covariate-adaptive randomization.

Keywords: Stratified permuted block randomization; Biased coin randomization; Cox model; Sensitivity analysis; Type I error rate; Power

Comparing MCP-MOD and Ordinal Linear Contrast Test in Dose Finding Clinical Trials: A Thorough Examination

Yaohua Zhang, Ning Li, **Naitee Ting**

Vice President, StatsVita, LLC

ABSTRACT

The MCPMod approach to design and analyze Phase II clinical trials was first introduced in the early 2000 era. It has successfully revolutionized the thinking and practice in Phase II clinical development of new drugs. However, in many situations, the assumptions behind the dose-response relationship were not realistic, and in other situations, models could be misleading. There is a need to simplify the understanding and practice of Phase II clinical development programs. If monotonic dose-response relationship can be assumed, then the ordinal linear contrast test (OLCT) approach introduced in 2017 can be considered as an improvement of MCPMod. One important contribution in OLCT is that it is easy to communicate with non-statisticians. This property largely improved the quality of teamwork in project teams and trial teams.

Keywords: Dose Finding, Ordinal Linear Contrast Test, MCP-Mod, Phase II Clinical Trial

Naitee Ting is a Fellow of American Statistical Association (ASA). He is currently Vice President of Veramed. Naitee is also an Adjunct Professor of Department of Statistics at University of Connecticut, Adjunct Professor of Department of Biostatistics at Columbia University. He joined Veramed in 2025. Before Veramed, Naitee has been with Boehringer Ingelheim Pharmaceuticals, Inc. (BI) for 15 years, and he was working at Pfizer Inc. for 22 years (1987-2009). Naitee received his Ph.D. in 1987 from Colorado State University (major in Statistics). He has an M.S. degree from Mississippi State University (1979, Statistics) and a B.S. degree from College of Chinese Culture (1976, Forestry) at Taipei, Taiwan.

Naitee published articles *in Technometrics, Statistics in Medicine, Drug Information Journal (Therapeutic Innovation and Regulatory Science), Journal of Statistical Planning and Inference, Journal of Biopharmaceutical Statistics, Biometrical Journal, Statistics and Probability Letters, Statistics in Biosciences, Statistics in Biopharmaceutical Research,* and *Journal of Statistical Computation and Simulation*. His book "*Dose Finding in Drug*

Development" was published in 2006 by Springer, and is considered as the leading reference in the field of dose response clinical trials. The book "*Fundamental Concepts for New Clinical Trialists*", co-authored with Scott Evans, was published by CRC in 2015. Another book "*Phase II Clinical Development of New Drugs*", co-authored with Chen, Ho, and Cappelleri was published in 2017 (Springer). Naitee has been an active member of both the ASA and the International Chinese Statistical Association (ICSA).

Back to Sessions List

Simultaneous Clustering and Estimation of Additive Shape Invariant Models for Recurrent Event Data

Zitong Zhang[1], **Shizhe Chen[1]**

Department of Statistics, University of California, Davis

ABSTRACT

Technological advancements have enabled the recording of spiking activities from large neuron ensembles, presenting an exciting yet challenging opportunity for statistical analysis. This project considers the challenges from a common type of neuroscience experiments, where randomized interventions are applied over the course of each trial. The objective is to identify groups of neurons with unique stimulation responses and estimate these responses. The observed data, however, comprise superpositions of neural responses to all stimuli, which is further complicated by varying response latencies across neurons. We introduce a novel additive shape invariant model that is capable of simultaneously accommodating multiple clusters, additive components, and unknown time-shifts. We establish conditions for the identifiability of model parameters, offering guidance for the design of future experiments. We examine the properties of the proposed algorithm through simulation studies, and apply the proposed method on neural data collected in mice.

Keywords: Point processes, shape invariant models, additive models, clustering

Multiview Manifold Learning for High-Dimensional and Noisy Data Analysis

Xiucai Ding[1], Chao Shen and Hau-Tieng Wu[2,*]

Department of Statistics, University of California, Davis

ABSTRACT

A longstanding challenge in data science is to effectively quantify systems of interest by integrating information from heterogeneous datasets, a problem known as multiview learning. In this talk, I will present recent advancements in this direction, focusing on novel algorithms based on convolutions of diffusion maps or kernel embeddings. Within the common manifold framework, the proposed algorithm can be interpreted through its spectral connection to limiting Laplacian or integral operators. Additionally, we demonstrate that the method is robust against high-dimensional noise via the analysis of the underlying kennel random matrices.

Keywords: Multiview Learning; Manifold Learning; Representation Learning; High-dimensional data analysis

Quantile Small Area Estimation via Singh-Maddala Mixed Model Prediction

Thuan Nguyen[1], Yuzi Liu[2], Haiqiang Ma[2], Xiaohui Liu[2] and Jiming Jiang[3,*]

[1]*OHSU/PSU School of Public Health, Oregon Health and Science University*

[2]*School of Statistics and Data Science, Jiangxi University of Finance and Economics*

[3]*Department of Statistics, University of California, Davis*

ABSTRACT

We develop methods of mixed model prediction (MMP) of quantiles of interest under the Singh-Maddala (SM) distribution for the outcome variable, which is widely used for financial and economic data. Such outcome data are often non-Gaussian, having skewed distributions. The traditional quantile mixed effects models have largely focused on estimating the fixed effects and variance components in the model. However, prediction of quantiles at subject-level, such as those associated with the small areas, are also of practical interest. We develop methods of MMP for subject-level quantiles of interest under the SM outcome distribution. Specifically, we develop the optimal prediction theory under a mixed effects SM model. The best predictor (BP) under the expected pinball loss (EPL) is the conditional quantile of the target interest. We then obtain the empirical BP (EBP) of the subject-level quantiles by replacing the unknown parameters involved in the BP by their maximum likelihood estimators. We establish asymptotic optimality of the EBP under the EPL measure. As a measure of uncertainty, we develop a second-order unbiased estimator of the EPL of the EBP. Empirical performances of the EBP and its EPL estimator are evaluated via simulation studies. An application to income inequality in China is discussed.

Keywords: Best prediction, Financial and economic data, Mixed-effect SM model, Subject-level quantiles

Interpretable Transformer Regression for Functional and Longitudinal Covariates

Yuan-Jung Cynthia Juang, Jane-Ling Wang*
Department of Statistics, University of California, Davis

ABSTRACT

Predicting scalar outcomes from functional data is challenging when measurements are sparse, irregular, and noisy, as in many scientific and clinical longitudinal studies. We propose IDAT, a dual-attention Transformer that operates directly on masked sampling schedules and avoids ad-hoc imputation. IDAT couples (i) time-point attention, which captures local and long-range temporal dynamics together with the response relationship nonparametrically, with (ii) inter-sample attention, which adaptively shares information across subjects with similar trajectories to stabilize estimation under sparsity. These pathways complement one another: time-point attention captures subject-specific dynamics, whereas inter-sample attention leverages population structure to ``borrow information'' from other subjects, echoing principles from random-effects model in longitudinal analysis. Under a random-effects framework that accounts for irregular sampling and measurement noise, we prove prediction-error bounds and show that IDAT consistently approaches the oracle solution. Across both simulations and real-world applications, IDAT achieves the best overall performance among 19 baselines. Only in the extremely dense case (> 80% observations) TabPFN (a recent method published in Nature) achieve a slight advantage, while IDAT still significantly outperforms all other baselines in this scenario. The learned attention weights are interpretable, revealing predictive time domains and potential clusters. In conclusion, IDAT, an end-to-end sparsity-aware Transformer architecture, offers improvements both in predictive accuracy and interpretability for scalar-on-function prediction.

Keywords: Dual-attention Transformer, interpretable attention, functional/longitudinal data, irregular sampling plan

Joint Mixed Membership Modeling of Multivariate Longitudinal and Survival Data

Yuyang He[1], **Xinyuan Song**[2], Kai Kang[3]

Department of Statistics, The Chinese University of Hong Kong

ABSTRACT

This study develops a novel joint mixed membership model for multivariate longitudinal AD-related biomarkers and time of AD diagnosis. Unlike conventional finite mixture models that assign each subject a single subgroup membership, the proposed model assigns partial membership across subgroups, allowing subjects to lie between two or more subgroups. This flexible structure enables individualized disease progression and facilitates the identification of clinically meaningful neurological statuses that are often elusive in current mixed-effects models. We employ a spline-based trajectory model to characterize complex and possibly nonlinear patterns of multiple longitudinal clinical markers. A Cox model is then used to examine the effects of time-variant risk factors on the hazard of developing AD. We develop a Bayesian method coupled with efficient Markov chain Monte Carlo sampling schemes to perform statistical inference. The proposed approach is assessed through extensive simulation studies and an application to the Alzheimer's Disease Neuroimaging Initiative study, demonstrating better performance in AD diagnosis compared to existing joint models.

Keywords: Mixed membership model; longitudinal data; MCMC methods; survival data

Building a Dose Toxo-Equivalence Model from a Bayesian Meta-Analysis of Published Clinical Trials

Elizabeth A Sigworth[1], Samuel M Rubinstein[2], Jeremy L Warner[3], Yong Chen[4], **Qingxia Chen[1]**

[1]*Department of Biostatistics, Vanderbilt University.*

[2]*Division of Hematology, University of North Carolina School of Medicine.*

[3]*Department of Medicine, Vanderbilt University School of Medicine.*

[4]*Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania.*

ABSTRACT

In clinical practice, medications are often interchanged in treatment protocols when a patient negatively reacts to their first line of therapy. Although switching between medications is common, clinicians often lack structured guidance when choosing the initial dose and frequency of a new medication, given the former with respect to the risk of adverse events. In this paper, we propose to establish this dose toxic-equivalence relationship using published clinical trial results with one or both drugs of interest via a Bayesian meta-analysis model that accounts for both within- and between-study variances. With the posterior parameter samples from this model, we compute median and 95% credible intervals for equivalent dose pairs of the two drugs that are predicted to produce equal rates of an adverse outcome, relying solely on study-level information. Via extensive simulations, we show that this approach approximates well the true dose toxo-equivalence relationship, considering different study designs, levels of between-study variance, and the inclusion/exclusion of no confounder/nonmodifier subject-level covariates in addition to study-level covariates. We compare the performance of this study-level meta-analysis estimate to the equivalent individual patient data meta-analysis model and find comparable bias and minimal efficiency loss in the study-level coefficients used in the dose toxo-equivalence relationship. Finally, we present the findings of our dose toxo-equivalence model applied to two chemotherapy drugs, based on data from 169 published clinical trials.

Keywords: Bayesian methods; dose toxo-equivalence; individual patient data meta-analysis; study-level meta-analysis.

Quasi Instrumental Variable Methods for Stable Hidden Confounding and Binary Outcome

Zhonghua Liu[1], Baoluo Sun[2], Ting Ye[3], David Richardson[4], Eric Tchetgen Tchetgen[5]

[1]*Department of Biostatistics, Columbia University*

[2]Department of Statistics and Data Science, National University of Singapore

[3]Department of Biostatistics, University of Washington, Seattle

[4]Department of Environmental and Occupational Health, University of California, Irvine

[5]Department of Statistics and Data Science, University of Pennsylvania

ABSTRACT

Instrumental variable (IV) methods are central to causal inference from observational data, particularly when a randomized experiment is not feasible. However, of the three conventional core IV identification conditions, only one, IV relevance, is empirically verifiable; often one or both of the other conditions, exclusion restriction and IV independence from unmeasured confounders, are unmet in real-world applications. These challenges are compounded when the outcome is binary, a setting for which robust IV methods remain underdeveloped. A fundamental contribution of this paper is the development of a general identification strategy justified under a structural equilibrium dynamic generative model of so-called stable confounding and a quasi-instrumental variable (QIV), i.e. a variable that is only assumed to be predictive of the outcome. Such a model implies (a) stability of confounding on the multiplicative scale, and (b) stability of the additive average treatment effect among the treated (ATT), across levels of that QIV. The former is all that is necessary to ensure a valid test of the causal null hypothesis; together those two conditions establish nonparametric identification and estimation of the conditional and marginal ATT. To address the statistical challenges posed by the need for boundedness in binary outcomes, we introduce a generalized odds product re-parametrization of the observed data distribution, and we develop both a principled maximum likelihood estimator and a triply robust semiparametric locally efficient estimator, which we evaluate through simulations and an empirical application to the UK Biobank.

Keywords: Binary outcome; Invalid instrument; Mendelian randomization; Semiparametric theory; Unmeasured confounding

Clustering-Informed Shared-Structure Variational Autoencoder for Missing Data Imputation in Large-Scale Healthcare Data

Yuan Chen[1], Yasin Khadem Charvadeh, Kenneth Seier, Katherine S. Panageas, Danielle Vaithilingam, Mithat Gönen, Yuan Chen[2,*]

Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center

ABSTRACT

Despite advancements in healthcare data management, missing data in electronic health records (EHR) and patient-reported outcomes remain a persistent challenge, limiting their usability in healthcare analytics. Conventional imputation methods often struggle to capture complex non-linear relationships, require extensive computation time, and are limited in addressing various types of missing data mechanisms. To overcome these challenges, we propose the clustering-informed shared-structure variational autoencoder (CISS-VAE), which utilizes the strengths of Bayesian neural networks. This model can effectively capture complex associations and accommodate various missing data mechanisms, including missing not at random (MNAR). We also develop iterative learning algorithms that further enhance missing data imputation accuracy while preventing overfitting. Comprehensive simulations demonstrate the superior accuracy of our model compared to traditional and contemporary methods. We apply our method to EHR data from early-stage breast cancer patients at Memorial Sloan Kettering Cancer Center, aiming to mitigate the impact of missing data and enhance health monitoring and analyses.

Keywords: Missing Data Imputation; Variational Autoencoder; Missing Not at Random; Electronic Health Records

Modeling Amplitude and Phase Variation of Multivariate Random Processes in Geodesic Spaces

Yaqing Chen, Kyum Kim

Department of Statistics, Rutgers University

ABSTRACT

For real-valued functional data, it is well known that failing to distinguish between amplitude variation and phase variation can distort subsequent statistical analysis, and extensive work has been devoted to developing time-warping methods to address this issue. However, much less is known about how to characterize and handle these two sources of variability when they co-exist in random processes taking values in general metric spaces, which lack inherent linear structure, particularly in the multivariate setting. In this paper, we formalize the concepts of amplitude and phase variation for multivariate random processes in geodesic spaces and propose a latent deformation model for jointly analyzing both types of variation. We establish the asymptotic convergence rates for the model estimators and demonstrate the versatility of the proposed method through simulation studies on positive semi definite (PSD) matrix-valued data and network-valued data, as well as a real-world application to annual sex-specific age-at-death distributions across countries, providing new insights into global longevity dynamics.

Keywords: distributional data; functional data; metric spaces; networks; PSD matrices; time warping

Transfer Learning for Functional Linear Regression

Xiaoyu Hu[1], **Zhenhua Lin[2]**

[1]*Xi'an JiaoTong University*

[2]*National University of Singapore*

ABSTRACT

We explore functional linear regression under posterior drift with transfer learning. Specifically, we investigate when and how auxiliary data can be leveraged to improve the estimation accuracy of the slope function in the target model when posterior drift occurs. We employ the approximated least square method together with a lasso penalty to construct an estimator that transfers beneficial knowledge from source data. Theoretical analysis indicates that our method avoids negative transfer under posterior drift, even when the contrast between slope functions is quite large. Specifically, the estimator is shown to perform at least as well as the classical estimator using only target data, and it enhances the learning of the target model when the source and target models are sufficiently similar. Furthermore, to address scenarios where covariate distributions may change, we propose an adaptive algorithm using aggregation techniques. This algorithm is robust against non-informative source samples and effectively prevents negative transfer. Simulation and real data examples are provided to demonstrate the effectiveness of the proposed algorithm.

Keywords: Transfer learning; functional data analysis; data fusion.

Inference for Dispersion and Curvature of Random Objects

Hans-Georg Müller*

Department of Statistics, University of California, Davis

ABSTRACT

A basic statistical task is to quantify statistical dispersion or spread. When one deals with random objects located in general metric spaces a challenge is the absence of vector space structure. A CLT for the joint distribution of Fréchet variance and metric variance, two measures of dispersion in geodesic spaces, reveals that the Alexandrov curvature of the geodesic space determines the relationship between these two dispersion measures, which generally do not coincide. This relation can be harnessed to infer the underlying curvature of the data, which results from properties of both the space and the probability measure that generates the random objects. The resulting test for curvature is supported by theory and aids in determining the intrinsic curvature of the (sub)space where the objects reside. Its finite sample properties are demonstrated for various data types and applications. This talk is based on joint work with Wookyeong Song, UC Davis.

Keywords: Metric Statistics; Fréchet Variance; Metric Variance; Intrinsic Curvature

Improving Interpretability in Machine Learning Using Confidence Intervals in ALE Plots

John R. Stevens[1] and Matthew Lister[1,2]

[1]*Department of Mathematics and Statistics, Utah State University*

[2]*Space Dynamics Laboratory, Utah State University*

ABSTRACT

Machine learning models that predict a response based on a collection of predictor variables usually do not provide simple numeric summaries of predictor effects and so are often referred to as black box models. Accumulated local effects (ALE) plots have been developed to allow visual interpretability of predictor effects in such black box models. We present R package AleCI, which improves the original ALE implementation by adding a bootstrapped confidence interval around each prediction showing the range where the true value of the predictor's effect should exist, for categorical as well as continuous predictors. AleCI is applicable across a variety of machine learning models and updates the graphing capabilities of the original implementation by using ggplot2.

Keywords: machine learning, statistical visualization, model interpretability

Guided Data Visualization via Random Forests and Manifold Learning

Kevin Moon

Utah State University

ABSTRACT

The manifold assumption has been used in many machine learning applications to combat the curse of dimensionality and to visualize the data structure. Most manifold learning methods are unsupervised and typically focus on preserving the dominant structure and variation in the data. In many cases, we wish to analyze the data in a supervised setting with respect to expert-provided data labels. Most supervised manifold learning methods exaggerate the separation between data points of different classes, distorting the true structure of the data. In this talk, I will present RF-PHATE, a supervised dimensionality reduction method that preserves the true structure of the variables that are relevant for the supervised task.

Keywords: data visualization; manifold learning; dimensionality reduction

AI-Assisted Data Visualization and Analytics

Kwan-Liu Ma

Computer Science, University of California at Davis

ABSTRACT

Today AI technology is transforming every field, including the domain of visualization. As a powerful tool for data-driven problem-solving, decision-making, and storytelling, visualization must be thoughtfully designed to ensure both effectiveness and performance. I will show how AI and machine learning can help enhance the visualization and analytical reasoning process, drawing from projects conducted by my research group. These technologies hold significant promise in facilitating critical steps of data analysis, guiding in the visualization design space, and optimizing visual outputs to improve user experience and performance.

Keywords: Machine learning; artificial intelligence; data analytics; visualization

iISOMAP: Nonlinear Dimensionality Reduction and Visualization for Interval-Valued Data via Geodesic Distance Preservation

Han-Ming Wu*

Department of Statistics, National Chengchi University

ABSTRACT

Dimensionality reduction for interval-valued data remains an active area of research within symbolic data analysis (SDA). While most existing approaches have primarily focused on adapting linear methods, such as principal component analysis (PCA), to interval data, this study introduces a novel nonlinear approach, namely interval ISOMAP (iISOMAP), which extends the isometric feature mapping (ISOMAP) technique to interval-valued data, explicitly focusing on geodesic distance preservation to uncover the intrinsic geometric structure of datasets. Unlike PCA, ISOMAP is a nonlinear dimensionality reduction (NLDR) method that reconstructs manifold structures by leveraging shortest-path distances on a neighborhood graph. The core innovation of iISOMAP lies in its use of interval multidimensional scaling (MDS) to accurately estimate and preserve geodesic distances between interval-valued data points. For visualization, the maximum covering area rectangle (MCAR) method is employed, projecting interval objects onto a two-dimensional NLDR subspace while maintaining their inherent uncertainty. We evaluate the performance of iISOMAP on both simulated and real-world datasets, comparing it against interval PCA and interval MDS to demonstrate its superior ability to capture nonlinear manifolds. Furthermore, we propose a novel aggregation method that summarizes a dataset's nonlinear structure into interval-valued representations, enhancing iISOMAP's applicability across a range of symbolic data analysis tasks.

Keywords: Interval multidimensional scaling; Isometric feature mapping; Symbolic data analysis

Bivariate Analysis of Distribution Functions Under Biased Sampling

Hsin-wen Chang[1], Shu-Hsiang Wang[1]

[1]*Institute of Statistical Science, Academia Sinica, Taipei, Taiwan*

ABSTRACT

We compare distribution functions among pairs of locations in their domains, in contrast to the typical approach of univariate comparison across individual locations. This bivariate approach is studied in the presence of sampling bias, which has been gaining attention in infectious disease studies that over-represent more symptomatic people. In cases with either known or unknown sampling bias, we introduce Anderson--Darling-type tests based on both the univariate and bivariate formulation. A simulation study shows the superior performance of the bivariate approach over the univariate one. We illustrate the proposed methods using real data on the distribution of the number of symptoms suggestive of COVID-19.

Keywords: Bootstrap; Empirical distribution function; Size bias; Two-sample test

Regression in 2-Wasserstein Distance

Li-Shan Huang[1], Zhezhen Jin[2], and Ting-Wei Hsu[1]

[1]*Institute of Statistics and Data Science, National Tsing Hua University, Hsinchu, TAIWAN*

[2]*Department of Biostatistics, Columbia University, New York, USA*

ABSTRACT

With the advancement of technology, there is a growing need for regression tools capable of modeling relationships in which the response variables are histograms or probability density functions, and the covariates lie in Euclidean space. In this work, we revisit classical linear regression and reformulate it from the perspective of minimizing 2-Wasserstein distances. The results lead to the development of 2-Wasserstein distance regression for histogram or density response objects with Euclidean predictors. The performance of the proposed methodology is evaluated through simulations, and its practical application is demonstrated through an analysis of mortality data.

Keywords: Best linear unbiased estimator; Gauss-Markov Theorem; Least squares

Debiased Inference for High-Dimensional Censored Quantile Regression

Yu Guo, **Tony Sit**

Department of Statistics and Data Science, The Chinese University of Hong Kong

ABSTRACT

This paper introduces a novel methodology for constructing confidence intervals in high-dimensional censored quantile regression, where the number of covariates may substantially exceed the sample size. Building upon the weighted loss function proposed by Wang and Wang (2009), we incorporate an L1 penalty to handle high dimensionality and apply a debiasing procedure to correct the inherent bias introduced by the LASSO estimator. The resulting debiased estimator is shown to be asymptotically normal, forming a solid foundation for valid statistical inference. Notably, our approach relaxes the conventional global linearity assumption to a local linearity condition near the quantile of interest, enhancing model flexibility and robustness—especially in the presence of heteroskedasticity or violations of global linear effects. Simulation studies demonstrate the superior performance of our method in terms of coverage accuracy and efficiency when constructing confidence intervals, compared to existing approaches. The practical utility of the method is further illustrated through an application to the lung cancer dataset of Shedden (2008), yielding the identification of potentially significant genes associated with survival outcomes.

Keywords: Conditional quantiles; Right-censored data; Post-selection inference; Debiasing

Data Integration in Survey Sampling and Official Statistics

Changbao Wu

Department of Statistics and Actuarial Science, University of Waterloo

ABSTRACT

We discuss issues with and techniques for combining information from multiple sources under settings involving probability samples, non-probability samples, multiple-frame survey samples, known population controls from a census or estimated population controls from other surveys. Two main principles, namely, *validity* and *efficiency*, for methodological developments are discussed under different scenarios. Inferential techniques are reviewed under three general frameworks for analysis of non-probability survey samples, namely, inverse probability weighting, mass imputation, and doubly robust estimation. Some recent research on using modern machine learning techniques for data integration will be briefly discussed.

Keywords: calibration; estimating equations; non-probability samples; variance estimation

Mask-Conditional Conformal Prediction: Valid Uncertainty for All Missing Data Mechanisms

Jiarong Fan[1,2], Juhyun Park[1,3], Thi Phuong Thuy Vo[1,3], Nicolas Brunel[1,3,4]

[1]LaMME, [2]University Paris Saclay, [3]ENSIIE, [4]Capgemini Invent

ABSTRACT

Conformal prediction (CP) offers a principled framework for uncertainty quantification, but it fails to guarantee coverage when faced with missing covariates. In addressing the heterogeneity induced by various missing patterns, Mask-Conditional Valid (MCV) Coverage has emerged as a more desirable property than Marginal Coverage. In this work, we adapt split CP to handle missing values by proposing a preimpute-mask-then-correct framework that can offer valid coverage. We show that our method provides guaranteed Marginal Coverage and Mask-Conditional Validity for general missing data mechanisms. A key component of our approach is a reweighted conformal prediction procedure that corrects the prediction sets after distributional imputation (multiple imputation) of the calibration dataset, making our method compatible with standard imputation pipelines. We derive two algorithms, and we show that they are approximately marginally valid and MCV. We evaluate them on synthetic and real-world datasets. It reduces significantly the width of prediction intervals w.r.t standard MCV methods, while maintaining the target guarantees.

Keywords: Conformal Prediction, Missing Data, Weighted Conformal Prediction, Uncertainty Estimation

Robust Conformal Prediction Using Privileged Information

Shai Feldman[1], Yaniv Romano[1,2]

[1]*Computer Science Department, Technion*

[2]*Electrical and Computer Engineering Department, Technion*

ABSTRACT

We develop a method to generate prediction sets with a guaranteed coverage rate that is robust to corruptions in the training data, such as missing or noisy variables. Our approach builds on conformal prediction, a powerful framework to construct prediction sets that are valid under the i.i.d assumption. Importantly, naively applying conformal prediction does not provide reliable predictions in this setting, due to the distribution shift induced by the corruptions. To account for the distribution shift, we assume access to privileged information (PI). The PI is formulated as additional features that explain the distribution shift, however, they are only available during training and absent at test time. We approach this problem by introducing a novel generalization of weighted conformal prediction and support our method with theoretical coverage guarantees. We further we analyze the robustness of our method to inaccuracies in the weights. Our analysis indicates that our proposal can still yield valid uncertainty estimates even when the weights are poorly estimated. Empirical experiments on both real and synthetic datasets indicate that our approach achieves a valid coverage rate and constructs more informative predictions compared to existing methods, which are not supported by theoretical guarantees.

Keywords: Conformal Prediction, Uncertainty Quantification, Distribution Shift, Corrupted Data, Privileged Information

Adaptive Coverage Policies in Conformal Prediction

Etienne Gauthier[1], Francis Bach[1], Michael I. Jordan[1,2]

[1]*Inria, Ecole Normale Supérieure, PSL Research University*

[2]*Departments of EECS and Statistics, University of California, Berkeley*

ABSTRACT

Traditional conformal prediction methods construct prediction sets such that the true label falls within the set with a user-specified coverage level. However, poorly chosen coverage levels can result in uninformative predictions, either producing overly conservative sets when the coverage level is too high, or empty sets when it is too low. Moreover, the fixed coverage level cannot adapt to the specific characteristics of each individual example, limiting the flexibility and efficiency of these methods. In this work, we leverage recent advances in e-values and post-hoc conformal inference, which allow the use of data-dependent coverage levels while maintaining valid statistical guarantees. We propose to optimize an adaptive coverage policy by training a neural network using a leave-one-out procedure on the calibration set, allowing the coverage level and the resulting prediction set size to vary with the difficulty of each individual example. We support our approach with theoretical coverage guarantees and demonstrate its practical benefits through a series of experiments.

Keywords: Conformal prediction; E-values; Uncertainty quantification; Machine learning

Towards a Rigorous Evaluation of RAG Systems: The Challenge of Due Diligence

Grégoire Martinon, Alexandra Lorenzo de Brionne, Jérôme Bohard, Antoine Lojou, Damien Hervault, **Nicolas Brunel**

Invent Lab, Capgemini Invent, France and ENSIIE, LaMME, Université Paris Saclay

ABSTRACT

The rise of generative AI has driven significant advancements in high-risk sectors like healthcare and finance. The Retrieval-Augmented Generation (RAG) architecture, combining language models (LLMs) with search engines, is particularly notable for its ability to generate responses from document corpora. Despite its potential, the reliability of RAG systems in critical contexts remains a concern, with issues such as hallucinations persisting. This study evaluates a RAG system used in due diligence for an investment fund. We propose a robust evaluation protocol combining human annotations and LLM-Judge annotations to identify system failures, like hallucinations, off-topic, failed citations, and abstentions. Inspired by the Prediction Powered Inference (PPI) method, we achieve precise performance measurements with statistical guarantees. We provide a comprehensive dataset for further analysis. Our contributions aim to enhance the reliability and scalability of RAG systems evaluation protocols in industrial applications. To ensure consistency in the online technical program, this abstract template must be used for submissions.

Keywords: LLM, RAG, hallucinations, LLM-as-a-Judge

Univariate Self-Starting Shiryaev (3S): A Bayesian Change Point Model for Online Monitoring of Short Runs

Konstantinos Bourazas[1], **Panagiotis Tsiamyrtzis[2,3]**

[1]*Department of Economics, Athens University of Economics and Business, Greece*

[2]*Department of Mechanical Engineering, Politecnico di Milano, Italy*

[3]*Department of Statistics, Athens University of Economics and Business, Greece*

ABSTRACT

The Shiryaev's change point methodology is a powerful Bayesian tool in detecting persistent parameter shifts. It has certain optimality properties when we have pre/post-change known parameter setups. In this work we will introduce a self-starting version of the Shiryaev's framework that could be employed in performing online change point detection in short production runs. Our proposal will utilize available prior information regarding the unknown parameters, breaking free from the phase I requirement and will introduce a more flexible prior for change-point parameter, compared to what standard Shiryaev employs. Apart from the on-line monitoring, our proposal will provide posterior inference for all the unknown parameters, including the change point. The modeling will guard in detecting persistent parameter shifts. A real data set will illustrate its use, while a simulation study will evaluate its performance against standard competitors.

Keywords: Bayesian Statistical Process Control and Monitoring, At Most Once Change (AMOC), Persistent Shifts, Phase I

Adaptive Sampling in Profile Monitoring Through Bandits

Chen Nan, Liu Peiyao*

Department of Industrial Systems Engineering and Management, National University of Singapore

ABSTRACT

Profile monitoring has found many applications in industrial problems. In this research, we consider the adaptive sampling strategies in profile monitoring. Instead of uniformly sample the profiles, we can adaptively choose the number and locations of the samples to increase the chance of anomaly detection. We use online learning approach to balance the trade off between exploration and exploitation in sampling to achieve good performance in detecting changes in different forms. Numerical studies have demonstrated the superior performance over conventional approach with uniform samplings.

Keywords: Profile monitoring, adaptive sampling, upper confidence bound

Multivariate Control Charts for Correlated Quality Variables of Different Types

Arthur B. Yeh

Department of Business Analytics, Economics and Information Systems

The Schmidthorst College of Business

Bowling Green State University, Bowling Green, Ohio, U.S.A.

ABSTRACT

The two major challenges, monitoring/detection and diagnostics, for multivariate control chart developments and applications have become more front and center in modern era, especially for non-manufacturing processes, as these processes often involve correlated variables of different types, continuous, count and categorical. Little attention has been paid to develop multivariate control charts for monitoring correlated variables of different types. Even fewer efforts have been devoted to developing diagnostic mechanisms for identifying out-of-control parameters under such a premise. In this talk, we will discuss how these two challenges present a unique opportunity to develop multivariate control charts which can not only monitor correlated variables of different types, but also provide instantaneous diagnostics of out-of-control parameters when an out-of-control signal is detected. The discussions focus on some recent works which tackle these challenges by adopting multiple testing procedures in developing multivariate control charts. Future research directions along the same line will also be discussed.

Keywords: Diagnostic mechanism; Multiple testing procedure; Multivariate control chart; Variables of different types

Privacy-Preserving LLM Alignment via Private Reward Modeling: A Holistic and Data-Efficient Framework

Young Hyun Cho[1], Will Wei Sun[2]

[1]*Department of Statistics, Purdue University*

[2]*Mitch Daniels School of Business, Purdue University*

ABSTRACT

Adapting Large Language Models (LLMs) to specific domains via preference alignment is essential for capturing nuanced human expectations, yet it introduces significant privacy risks when sensitive data is involved. While Differential Privacy (DP) is the standard solution, applying it to the preference fine-tuning stage presents a critical dilemma. Standard multi-stage pipelines necessitate inefficient data partitioning that limits utility, whereas applying DP directly to methods like Direct Preference Optimization (DPO) suffers from severe gradient instability. Furthermore, common Label-DP approaches fail to protect user prompts, leaving sensitive interaction details exposed.

In this work, we introduce a framework that decouples the privacy mechanism from policy optimization to achieve stability and holistic, Tuple-level DP. Our approach learns a Private Reward Model using the entire dataset—resolving partitioning inefficiencies—and derives the final policy via a deterministic post-processing step, thereby circumventing unstable gradient updates. This guarantees DP for the entire interaction tuple (prompt, response, and label), offering significantly stronger protection than Label-DP. Theoretically, we establish sample complexity bounds matching the non-private rate up to an additive privacy cost. Empirically, our method significantly outper- forms private DPO and PPO baselines on the Gemma-2b-it task, highlighting improved stability and privacy–utility trade-offs relative to contemporary baselines.

Keywords: Large Language Models, Reinforcement Learning with Human Feedback, Differential Privacy, Reward Modeling, Sample Complexity

Bridging Spatial Transcriptomics and Histopathology through AI

Wei Chen[1,*], Chongyue Zhao[1], Tianhao Liu[1]

(Author order follows that of the original publication)

Department of Pediatrics, University of Pittsburgh,

UPMC Children's Hospital of Pittsburgh, Pittsburgh, PA, USA

ABSTRACT

Spatial transcriptomics (ST) technologies from multicellular platforms such as Visium and Curio to subcellular platforms including Visium HD and Stereo-seq remain limited in resolving gene expression at true single-cell resolution. Existing computational methods operate largely at the spot level and cannot accurately reconstruct full transcriptomes for individual cells across ST resolutions. We introduce a unified framework that integrates high-resolution histology images with spot-level ST data using a Vision Transformer model and contrastive learning. It reconstructs single-cell gene expression profiles from both multicellular and subcellular platforms. Using in-house mouse lung datasets generated on Visium, Visium HD, and Stereo-seq, along with public datasets, our method consistently improves biological signal recovery: (1) at the tissue level, it delineates structural domains and immune regions (2) at the cellular level, it identifies major immune populations, resolves CAF and macrophage subtypes including rare $SPP1^{+}$ macrophages, and detects tertiary lymphoid structures in colorectal cancer, and (3) at the molecular level, it enhances cell-type separation and differential expression accuracy. Beyond ST reconstruction, our method further extends to histology-only cell annotation, leveraging learned multimodal representations to classify individual cells on routine H&E images, including those from complex disease tissues. In addition, a user-friendly interactive interface enables real-time visualization, expert refinement, and scalable annotation, supporting broad translational applications in spatial genomics and digital pathology.

Keywords: Spatial Transcriptomics; Vision Transformer; Histopathology; scRNA-seq.

Fair Graph Learning Without Complete Demographics

Zichong Wang[1], **Fang Liu**[*2], Shimei Pan[3], Jun Liu[4], Fahad Saeed[1], Meikang Qiu[5], Wenbin Zhang[1]

[1]*Florida International University, FL, USA*

[2]*University of Notre Dame, IN, USA*

[3]*University of Maryland Baltimore County, MD, USA*

[4]*Northeastern University, MA, USA*

[5]*Augusta University, GA, USA*

ABSTRACT

Graph Neural Networks (GNNs) have excelled in diverse applications due to their outstanding predictive performance, yet they often overlook fairness considerations, prompting numerous recent efforts to address this societal concern. However, most fair GNNs assume complete demographics by design, which is impractical in most real-world socially sensitive applications due to privacy, legal, or regulatory restrictions. For example, the Consumer Financial Protection Bureau mandates that creditors ensure fairness without requesting or collecting information about an applicant's race, religion, nationality, sex, or other demographics. We propose fairGNN-WOD, a first-of-its-kind framework that considers mitigating unfairness in graph learning without using demographic information. We analyze bias in node representations and establish the relationship between utility and fairness objectives. Experiments on three real-world graph datasets illustrate that fairGNN-WOD outperforms state-of-the-art baselines in achieving fairness but also maintains comparable prediction performance.

Keywords: AI Ethics, Trust, Fairness, Graph Neural Networks

Super Learner for Survival Prediction in Case-Cohort and Generalized Case-Cohort Studies

Jianwen Cai, Haolin Li, Haibo Zhou, David Couper

Department of Biostatistics, University of North Carolina at Chapel Hill

ABSTRACT

The case-cohort study design is often used in modern epidemiological studies of rare diseases, as it can achieve similar efficiency as a much larger cohort study with a fraction of the cost. Previous work focused on parameter estimation for case-cohort studies based on a particular statistical model, but few discussed the survival prediction problem under such type of design. In this article, we propose a super learner algorithm for survival prediction in case-cohort studies. We further extend our proposed algorithm to generalized case-cohort studies. The proposed super learner algorithm is shown to have asymptotic model selection consistency as well as uniform consistency. We also demonstrate our algorithm has satisfactory finite sample performances. Simulation studies suggest that the proposed super learners trained by data from case-cohort and generalized case-cohort studies have better prediction accuracy than the ones trained by data from the simple random sampling design with the same sample sizes. Finally, we apply the proposed method to analyze a generalized case-cohort study conducted as part of the Atherosclerosis Risk in Communities (ARIC) Study.

Keywords: Cost-Efficient Design; Ensemble Learning; Epidemiological Studies; Survival Analysis

Efficient Case-Cohort Design Using Balanced Sampling

Kaeum Choi[1], **Sangwook Kang[2]**

[1]*Department of Biostatistics and Bioinformatics, Emory University, Georgia, USA*

[2]*Department of Applied Statistics, Yonsei University, Seoul, Republic of Korea*

ABSTRACT

The case-cohort design is a cost-efficient two-phase design for analyzing survival data when key risk factors are expensive to assess and the event rate is low. Traditionally, subcohorts are selected via simple random sampling, which might not fully utilize available information. In this talk, we introduce an efficient sampling design using balanced sampling for subcohort selection within the case-cohort design. A notable benefit of employing balanced sampling is the automatic calibration of auxiliary variables available for the entire cohort. Under a Cox model, it has been demonstrated that the calibration of sampling weights, utilizing auxiliary variables highly correlated with the main risk factor, significantly enhances the efficiency of regression coefficient estimators. Extensive simulation experiments show the reduced variabilities under the proposed approach in comparison to those under both simple random sampling. The proposed design and estimation procedure are further illustrated using the well-established National Wilms Tumor Study dataset.

Keywords: Calibration; cohort sampling; Cox model; sampling weights; survival analysis

Improving Efficiency of Risk Prediction with Subsampled Cohort Data

Yei Eun Shin[1]

[1]*Department of Statistics, Seoul National University*

ABSTRACT

In large cohort studies, subsampling designs such as case-cohort and nested case-control sampling are widely used to improve efficiency and reduce costs. However, these designs introduce methodological challenges for survival analysis, particularly when estimating absolute risk, considering competing events, or handling time-varying covariates. This talk introduces recent methods that aim to improve the performance of survival analysis under such designs. The first part presents approaches for improving the estimation and validation of risk prediction models including absolute risks, proportional and additive hazards models when risk factors are not fully available in a cohort. For competing risk analysis with event-specific subsamples, a proportional risk model provides a simple and statistically efficient way in a joint framework. Additional topics include estimating transition probabilities in multi-state models and applying landmarking methods when only limited subsampled data are available. Together, these methods illustrate how targeted use of subsampled data via influence functions can support efficient and flexible survival analysis, even when full cohort data are not available.

Keywords: influence function; risk prediction model; two-phase sampling designs; weight calibration

Semiparametric Regression Analysis of Case-Cohort Studies with Multiple Interval-Censored Disease Outcomes

Haibo Zhou[1], Qinging Zhou[2], Jianwen Cai[1]

[1]*University of North Carolina at Chapel Hill, USA*

[2]*University of North Carolina at Charlotte, USA*

ABSTRACT

In this work, we formulate the case-cohort design with multiple interval-censored disease outcomes and generalize it to nonrare diseases where only a portion of diseased subjects are sampled. We develop a marginal sieve weighted likelihood approach, which assumes that the failure times marginally follow the proportional hazards model. We consider two types of weights to account for the sampling bias, and adopt a sieve method with Bernstein polynomials to handle the unknown baseline functions. We employ a weighted bootstrap procedure to obtain a variance estimate that is robust to the dependence structure between failure times. The proposed method is examined via simulation studies and illustrated with a dataset on incident diabetes and hypertension from the Atherosclerosis Risk in Communities study.

Keywords: case-cohort design, robust inference, sieve estimation, survival analysis

Choice of Metric and the Effect of Scan Length for Reliability in Resting-State fMRI

Yu Huang[1], Seonjoo Lee[2], Philip Reiss[3], and **R. Todd Ogden**[2]

[1]*Department of Population Science, New York University*

[2]*Department of Biostatistics, Columbia University*

[3]*Department of Statistics, University of Haifa*

ABSTRACT

Resting-state fMRI (rs-fMRI) studies are often used to study functional connectivity, i.e., the communication between R distinct regions when no specific task is being performed. Such data are often presented as R x R covariance matrices, each of which can be regarded as being a point on a manifold. We discuss measures of reliability for such data that are based on metrics that respect the manifold structure of the space. Applying these concepts, we explore aspects of design and data collection for such studies including region selection, scan length, and time interval between scans. We illustrate these concepts through application to rs-fMRI data from the Midnight Scanning Club dataset.

Keywords: resting-state fMRI; correlation matrices; object-valued data, Riemannian metric

Clarifying and Extending Permutation Tests on Brain Map Correspondence Through Mixed-Effects Modeling

Qiaochu Wang[1], Lingyi Peng[2], Thomas E. Nichols[3], Xu Zou[1], Yaotian Wang[4], Jie He[1], Yuexin Zhang[1], Dana L. Tudorascu[2], Diego Szczupak[5], Lauren Schaeffer[5], Emily S. Rothwell[5], Dieckhaus, Laurel[5], Stacey J. Sukoff Rizzo[5], Gregory W. Carter[6], Afonso C. Silva[5], and **Tingting Zhang[1]**

[1]*Department of Statistics, University of Pittsburgh*

[2]*Department of Biostatistics and Health Data Science, University of Pittsburgh*

[3]*Oxford Big Data Institute, University of Oxford*

[4]*Department of Biostatistics and Bioinformatics, Emory University*

[5]*Department of Neurobiology, University of Pittsburgh*

[6]*The Jackson Laboratory*

ABSTRACT

Permutation-based methods such as the spin test, BrainSMASH, and the SPICE test are widely used to assess correspondence between spatially distributed brain maps while accounting for their spatial autocorrelation. However, these methods define and evaluate correspondence in fundamentally different ways, making their results difficult to compare or interpret jointly. We address these limitations by introducing a two-factor mixed-effects modeling framework that decomposes brain map variability into inter-subject and spatial components. This formulation characterizes distinct sources of variability in brain maps and reveals how different permutation tests target correspondence related to the different components. Within this framework, we further show the analytical expressions of the null distributions of the spin test, BrainSMASH, and SPICE in terms of model parameters. This unified framework clarifies the fundamental distinctions among permutation tests, reveals their implicit assumptions, and provides a principled way to compare and interpet their results across diverse correspondence scenarios. Beyond clarifying existing methods, the modeling framework naturally motivates a bootstrap-based inference method for jointly quantifying multiple compositions of correspondence. Using simulations and empirical analyses of structural (cortical thickness vs. sulcal depth) and functional (language vs. motor contrast) brain maps, we demonstrate that the bootstrap-based method offers competitive type I error control, much higher statistical power, and richer insights into the compositions and sources of variations that give rise to brain map correspondence.

Keywords: Brain map correspondence; Mixed-effects models; Permutation tests; Bootstrap

Threshold Spatial Attention Transformer for Efficient Image Generation

Yuhan Geng, Wei Hao, **Jian Kang***

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, United States

ABTRACT

We propose the threshold spatial attention transformer (TSAT), a novel model for efficient and high-quality image generation, with a focus on medical imaging applications. Existing models such as Generative Adversarial Networks (GANs) and Diffusion Probabilistic Models (DDPMs) often face challenges related to data efficiency, computational demands, and reliance on extensive labeled datasets. The proposed TSAT model addresses these limitations through a block-wise feature sampling mechanism integrated with special transformer architectures. Using an encoder-decoder framework, the mode effectively reduces the token dimensionality while preserving essential spatial and contextual information, enabling accurate image reconstruction and synthesis. Key innovations include low-rank approximation, spatial attention kernels, and nested thresholding techniques, which collectively improve computational efficiency. TSAT supports both supervised and semi-supervised training paradigms, demonstrating flexibility across different dataset sizes and labeling conditions. Numeric experiments on state-of-the-art computer vision datasets and the fMRI activation maps in the Human Brain Connectome (HCP) study highlight the superior performance of the TSAT, while requiring significantly smaller training sample size.

Keywords: Generative model, Neuroimaging, Deep neural networks

Matching for Causal Inference

Fang Han

Department of Statistics, University of Washington, Seattle, WA

ABSTRACT

In two landmark Econometrica papers, Abadie and Imbens established that the nearest neighbor (NN) matching estimator for the average treatment effect is asymptotically normal when using a fixed number of NNs, but it remains semiparametrically inefficient and bootstrap inconsistent. In this talk, I will demonstrate that these limitations can be overcome by simply allowing the number of NNs to grow with the sample size. Under this modification, the NN matching estimator becomes asymptotically normal, doubly robust, semiparametrically efficient, and bootstrap consistent. These results are published in Econometrica and other leading journals.

Keywords: graph-based statistics, graph central limit theorem, imputation, density ratio

Synthetic Nearest Neighbours: Extending Synthetic Controls for Matrix Completion with Missing Not at Random Data

Anish Agarwal[1], Munther Dahleh[2], Devavrat Shah[2], **Dennis Shen[3]**

[1]*Columbia University*

[2]*MIT*

[3]*USC*

ABSTRACT

We develop a causal framework for matrix completion under missing not at random (MNAR) data. Drawing on the method of synthetic controls from the econometric panel data literature, our approach relaxes two core assumptions that underlie standard MNAR matrix completion models: positivity (every entry is observed with positive probability) and independence (observations are independent across entries). Unlike traditional panel data models that often rely on rigid block-sparsity patterns, our framework accommodates flexible and heterogeneous observation structures commonly encountered in matrix completion problems. To operationalize our framework, we propose synthetic nearest neighbors (SNN), a novel algorithm that blends elements of K-nearest neighbors with synthetic controls. Under suitable assumptions on the underlying matrix and observed sparsity pattern, we prove that SNN achieves entrywise mean-squared error convergence for estimating the mean matrix, attaining a near-parametric rate. We further extend our analysis to heteroskedastic variance estimation, establishing that SNN attains entrywise mean-squared error convergence under bounded noise and asymptotic unbiasedness under general sub-Gaussian noise. Simulations studies corroborate our theoretical findings and demonstrate the robustness of SNN across a range of MNAR scenarios.

Keywords: Panel data, heteroskedastic variance estimation, causal inference

A Pleiotropy-Free Bayesian Model for Two- Sample Summary-Data Mendelian Randomization with Binary Outcomes

An-Shun Tai

Institute of Statistics and Data Science, National Tsing Hua University, Taiwan

ABSTRACT

Mendelian randomization (MR) is a powerful approach for estimating causal effects in the presence of unmeasured confounding. With the rapidly increasing sample sizes of genome-wide association studies, MR analyses based on twosample summary data across a broad range of phenotypes have become increasingly common. However, traditional summary-data MR techniques rely on the exclusion restriction, an assumption often violated due to pleiotropy. Although numerous methods have been proposed to address this issue, existing approaches still struggle to provide accurate causal effect estimates for binary outcomes. In this study, we derive the exact pleiotropy-induced bias under the counterfactual framework without requiring additional modeling assumptions. Building on this insight, we develop a pleiotropy-free Bayesian MR model for summary data with binary outcomes using a spike-and-slab prior to accommodate invalid instrumental variables. Through extensive simulation studies under diverse scenarios, we demonstrate the superior performance of our approach and offer a comprehensive comparison against current state-of-the-art methods.

Keywords: Mendelian Randomization; Causal Inference; Pleiotropy Bias; Bayesian Modeling; Two-Sample Summary Data.

Survival Data Analysis Using Average Hazard with Survival Weight

Hajime Uno, PhD

Dana-Farber Cancer Institute

Harvard Medical School

ABSTRACT

The traditional Cox hazard ratio has long been used to summarize treatment effects in time-to-event analyses, but its well-known limitations have led researchers to explore alternative measures. One such alternative, the average hazard with survival weight (AH), provides a person-time incidence rate that remains unaffected by nuisance random censoring, offering a more interpretable and robust summary of treatment effects. In this talk, we discuss this approach, including two-sample comparisons, regression analysis, and stratified analysis. By offering a fresh perspective on survival data analysis, the AH approach provides an impactful alternative to the Cox hazard ratio approach, helping researchers better capture and communicate the magnitude of treatment effects.

Keywords: Censoring, Cox regression, Hazard, Incidence Rate

Tools for Randomized Clinical Trials Using Restricted Mean Survival Time and Average Hazard

Miki Horiguchi

Dana-Farber Cancer Institute/Harvard T.H. Chan School of Public Health

ABSTRACT

In randomized clinical trials with time-to-event outcomes, the log-rank test based on Cox's proportional hazards (PH) model is commonly used for statistical comparisons, with the hazard ratio (HR) reported as the summary measure of treatment effect. However, the limitations of this traditional approach have been widely discussed. Alternative methods, such as restricted mean survival time (RMST) and average hazard with survival weight (AH), are gaining attention to address the limitations and providing more robust and interpretable quantitative information on treatment effects. While they have received attention, practical considerations for trial design using RMST or AH, particularly in determining analysis timing, remain understudied. We aim to fill these gaps by presenting methodological considerations and tools for identifying analysis timing, aiming to facilitate broader adoption of these alternative methods in practice.

Keywords: analysis timing, data monitoring, sample size calculation

Time-to-Event Analysis with Treatment Switches in Clinical Trials

Jie Chen

Taimei Intelligence Biopharma R&D

ABSTRACT

Patients in clinical trials often switch treatments during the study, e.g., switching from the originally assigned control treatment to the new treatment if the patients show disease progression and the new treatment is promising, or switching from the new treatment to the control treatment if the patients cannot tolerate the new treatment and the control treatment is a standard of care. The commonly encountered reasons for treatment switching encompass intolerability (e.g., serious adverse events), lack of efficacy (e.g., disease progression), decision or preference by patient and/or treating physician. Treatment switching in clinical trials generally satisfies ethical requirements but can complicate the estimation of treatment effects as measured by clinical outcomes, e.g., time-to-event endpoints, for the investigational drug after switching occurs. For example, the overall survival (OS) of switched participants reflects the combined effects of both pre- and post-switch treatments. Consequently, estimating the true survival benefit of the investigational drug may require adjustment for the impact of the post-crossover therapy. This talk will discuss the common scenarios for treatment switching, strategies of handling them, and statistical methods for estimating treatment effects as measured by time-to-event endpoints when treatment switching occurs. Real-world examples are provided to illustrate how these methods can be used in clinical studies.

Keywords: Estimand; causal inference; counterfactual methods; non-counterfactual methods; informative switching

Missing Data Strategies for Generalized Pairwise Comparisons in Randomized Clinical Trials

Ying Lu[1], Ruben P.A. van Eijk[2]

[1]*Stanford University School of Medicine, USA*

[2]*University Medical Center Utrecht, The Netherlands*

ABSTRACT

Generalized pairwise comparisons are increasingly being used in randomized clinical trials. This study evaluates how missing data strategies impact the operating characteristics of endpoints combining overall survival and a longitudinal outcome via a simulation study. Conditional longitudinal and survival data were generated based on the natural history of amyotrophic lateral sclerosis. Simulation scenarios varied in censoring rates, extent of missing longitudinal data, differential attrition between treatment arms, and the dependency of missingness on disease severity. All methods were unbiased and maintained nominal Type I error when censoring and missingness were balanced across treatment arms. Under imbalance, however, Type I error increased – reaching up to 0.378 for some strategies. The last common visit strategy was the only approach that consistently preserved nominal error rates (0.025 ± 0.002). Regarding statistical power, all methods exhibited a loss of precision and increased bias under the alternative hypothesis as missingness increased. Multiple imputation partially recovered power and reduced bias but inflated the Type I error rate in scenarios with differential attrition. As generalized pairwise comparisons are increasingly used in pivotal clinical trials – and therefore in regulatory contexts – our findings provide practical guidance for selecting a robust primary analysis strategy and minimizing bias arising from incomplete data.

Keywords: generalized pairwise comparisons; win statistics; hierarchical endpoint; combined assessment of function and survival; amyotrophic lateral sclerosis; clinical trials

December 20 (Saturday):

Parallel Sessions [8:40 – 10:20]:

20a1 - Machine Learning / AI

20a2 - Recent Advancements in Design of Experiments

20a3 - Causal Inference

20a4 - Bayesian Adaptive Designs for Oncology Dose Optimization

20a5 - Recent Advances in Network and Tensor Data Analysis

20a6 - Advances in Statistical Modelling and Inference for High-dimensional and Functional Data

20a7 - Recent Advances in Clinical Trials

20a8 - New Advances in Biostatistics and Bioinformatics

20a9 - Innovations in Statistical Methodology for Complex Data: Spatial Omics, Stochastic Optimization, Network Integration, and Microscopy Image Analysis

20a10 - Statistical and Algorithmic Foundations of Diffusion Models

Parallel Sessions [10:40 – 12:20]:

20b1 - Modern Functional Data Methods with Applications to Environmental Studies

20b2 - Subsampling

20b3 - Recent Advances in Time Series Analysis

20b5 - Recent Developments in Survival Analysis and Deep Learning

20b6 - Recent Advancements in Deep Learning and Graphical Models, and Model Selection

20b7 - Applied Probability

20b8 - Regulatory Advances in Clinical Trials

December 20 (Saturday):

Parallel Sessions [13:20 – 15:00]:

20c1 - New fronts on Machine learning

20c2 - Modern Advances in Learning, Robust Modeling, and Structure for High-Dimensional Data

20c3 - Advances in Financial Econometrics and Network Modeling

20c4 - Recent Advances in Data Science

20c5 - Lead Science and Clinical Research – Career Panel Discussion

20c6 - Flexible Inference: Nonparametrics, Causality, and Large Language Models

20c7 - Recent Advances in Network Data Analysis

20c8 - Complex Data Analysis in Environmental and Health Studies

20c9 - Statistical Analyses Methods for Integrating Multi-Omics Data with Application to Personalized Medication

20c10 - Causal Inference for Complex Designs

Parallel Sessions [15:20 – 17:00]:

20d1 - Causal Inference: Episode II

20d2 - Frontiers in Knowledge Creation and Discovery

20d3 - Advances in Stratified and Error-Prone Data Analysis

20d4 -

20d5 - Recent Advances in Statistical Learning for Complex Data Structures

20d6 - New Advances in Machine Learning and AI

20d7 - Innovations in Machine Learning for Financial Data Analysis

20d8 - Understanding the Biological Heterogeneity of Complex Traits through Omics Data

20d9 - Innovations in Survival Analysis and Clinical Trials for Biomedical Research

Modeling and Predicting Single-Cell Multi-Gene Perturbation Responses

Gefei Wang, Tianyu Liu, Jia Zhao, Youshu Cheng, **Hongyu Zhao***

Department of Biostatistics, Yale University

ABSTRACT

Understanding cellular responses to genetic perturbations is essential for deciphering gene regulation and phenotype formation. While high-throughput single-cell RNA-sequencing has facilitated detailed profiling of heterogeneous transcriptional responses to perturbations at the single-cell level, there remains a pressing need for computational models that can decode the mechanisms driving these responses and accurately predict outcomes to prioritize target genes for experimental design. This presentation introduces a deep generative learning framework designed to model and predict single-cell transcriptional responses to genetic perturbations, including single-gene and combinatorial multi-gene perturbations. The method effectively integrates prior biological knowledge and disentangles basal cell states from perturbation-specific salient representations by leveraging gene embeddings derived from large language models. Through comprehensive evaluations on multiple single-cell CRISPR Perturb-seq datasets, the approach outperformed state-of-the-art methods in predicting perturbation outcomes, achieving higher prediction accuracy. Notably, it demonstrated robust generalization to unseen target genes and perturbations, and its predictions captured both average expression changes and the heterogeneity of single-cell responses. Furthermore, the predictions enable diverse downstream analyses, including identifying differentially expressed genes and exploring genetic interactions, demonstrating its utility and versatility. This is joint work with Gefei Wang, Tianyu Liu, Jia Zhao, and Youshu Cheng.

Keywords: Statistical genetics, foundation models, embedding, biostatistics, genomics

Transcriptomic Analysis and Image-Based Deep Learning Prognostic Model for Lung Adenocarcinoma

Yang-Ming Yeh[1], Yahsuan Chang[2], Liang-Yin Tao[3], **Hsuanyu Chen[1]**

[1]*Institute of Statistical Science, Academia Sinica, Taipei City, Taiwan,*

[2]*National Health Research Institutes - Zhunan Campus, Zhunan Township, Taiwan*

[3]*Department of Statistics, University of California, Davis, Davis, CA, United States*

ABSTRACT

Background

Therapies such as EGFR tyrosine kinase inhibitors (TKIs) and immune checkpoint inhibitors (ICIs) have improved outcomes for EGFR-mutant and wild-type lung adenocarcinoma patients, respectively. However, drug resistance and limited survival remain significant challenges. This study proposes a novel prognosis prediction model using whole-transcriptome data integrated with an image-based deep learning approach.

Methods

Four cohorts were analyzed: one training cohort (RNA-seq, n=391) and three independent validation cohorts (TCGA RNA-seq, n=394; GSE68465, n=443; GSE13213, n=117). RNA-seq data were converted into images using pixel-encoding strategies to preserve transcriptomic information. A convolutional neural network (CNN) model was trained on all gene expression data without feature selection. To mitigate overfitting, the training cohort was split into training (n=259), testing (n=62), and validation (n=70) subsets. The CNN was trained on the training subset, validated on the testing and validation subsets, and independently evaluated on the three external validation cohorts.

Results

RNA-seq data were successfully transformed into images, enabling effective CNN model training. The model stratified patients into high- and low-risk groups. In the training cohort, high-risk patients exhibited significantly shorter overall survival (OS; $P < 0.01$). Similar findings were observed in the external validation cohorts: TCGA ($P = 0.001$), GSE68465 ($P < 0.0001$), and GSE13213 ($P = 0.003$). Prediction accuracies were 0.71, 0.81, and 0.92 for the training, testing, and validation subsets, respectively, with an average accuracy of 0.70 across the external cohorts.

Conclusions

This study presents an innovative image-based deep learning approach for analyzing whole-transcriptome data without requiring differential gene selection. By capturing comprehensive transcriptomic information, this method offers potential for enhanced prognostic modeling and molecularly guided lung cancer treatments.

Back to Sessions List

Modeling the Impact of Personal Genome Variation on Molecular Phenotypes

Nilah Ioannidis

UC Berkeley; UCSC; CZ Biohub

ABSTRACT

Understanding inter-individual variation in molecular, cellular, and other clinically-relevant phenotypes is an important challenge in precision medicine. Sequence-based genomic deep learning models that predict gene expression and other molecular phenotypes directly from DNA sequence can be applied in silico to sequences containing any combination of rare or common genetic variants, with great potential to predict the genetic contribution to variation in such phenotypes. However, despite success in explaining variation in molecular phenotypes across the genome and across a variety of cell types, we and others recently found that current sequence-based genomic deep learning models have limited ability to explain variation in gene expression across different individuals based on their personal genome sequences. I will discuss our work to characterize the cross-individual performance of such models on gene expression and other molecular phenotypes, with resulting insights into their understanding of regulatory variation. I will also discuss our recent efforts to develop models with improved understanding of variation across individuals using several strategies, such as incorporating personal genome and transcriptome data during model training and using a hierarchical approach to first model more locally-regulated phenotypes such as chromatin accessibility.

Keywords: Genomics; Machine Learning; Deep Learning.

An AI System to Help Scientists Write Expert-Level Empirical Software

Eser Aygün[1,*], Anastasiya Belyaeva[2,*], Gheorghe Comanici[1,*], **Dr. Marc A. Coram[2,*]**, Hao Cui[2,*], Jake Garrison[3,*], Renee Johnston[2,*], Anton Kast[2,*], Cory Y. McLean[2,*], Peter Norgaard[2,*], Zahra Shamsi[2,*], David Smalling[1,*], James Thompson[2,*], Subhashini Venugopalan[2,*], Brian P. Williams[2,*], Chujun He[2,4,**], Sarah Martinson[2,5,**], Martyna Plomecka[2,6,**], Lai Wei[2], Yuchen Zhou[2], Qian-Ze Zhu[2,5,**], Matthew Abraham[2], Erica Brand[2], Anna Bulanova[1], Jeffrey A. Cardille[2,7], Chris Co[2], Scott Ellsworth[2], Grace Joseph[2], Malcolm Kane[2], Ryan Krueger[2,5,**], Johan Kartiwa[2], Dan Liebling[2], Jan-Matthis Lueckmann[2], Paul Raccuglia[2], Xuefei (Julie) Wang[2,8,**], Katherine Chou[2], James Manyika[2], Yossi Matias[2], John C. Platt[2], Lizzie Dorfman[2], Shibl Mourad[1,‡] and Michael P. Brenner[2,5,‡]

Google Research, Google Zurich

ABSTRACT

The cycle of scientific discovery is frequently bottlenecked by the slow, manual creation of software to support computational experiments. To address this, we present an AI system that creates expert-level scientific software whose goal is to maximize a quality metric. The system uses a Large Language Model (LLM) and Tree Search (TS) to systematically improve the quality metric and intelligently navigate the large space of possible solutions. The system achieves expert-level results when it explores and integrates complex research ideas from external sources. The effectiveness of tree search is demonstrated across a wide range of benchmarks. In bioinformatics, it discovered 40 novel methods for single-cell data analysis that outperformed the top human-developed methods on a public leaderboard. In epidemiology, it generated 14 models that outperformed the CDC ensemble and all other individual models for forecasting COVID-19 hospitalizations. Our method also produced state-of-the-art software for geospatial analysis, neural activity prediction in zebrafish, time series forecasting and numerical solution of integrals. By devising and implementing novel solutions to diverse tasks, the system represents a significant step towards accelerating scientific progress.

Keywords: empiricial software, LLM-based data analysis, forecasting, batch-normalization

[1]Google DeepMind, [2]Google Research, [3]Google Platforms and Devices, [4]Massachusetts Institute of Technology, [5]School of Engineering and Applied Sciences, Harvard University,

[6]Google Cloud, [7]Faculty of Agricultural and Environmental Sciences, McGill University, [8]California Institute of Technology

*Equal contribution in alphabetical order

** Carried out as part of a student researchership at Google Research

‡ To whom correspondence should be addressed: shibl@google.com, mbrenner@google.com

Back to Sessions List

Optimal Designs for Network Experimentation with Unstructured Treatments

Ming-Chung Chang[1], **Jing-Wen Huang[2]**, Frederick Kin Hing Phoa[1]

[1]*Institute of Statistical Science, Academia Sinica, Taipei City, Taiwan,*

[2]*Institute of Statistics, National Tsing Hua University, Hsinchu, Taiwan*

ABSTRACT

Experiments involving connected units are prevalent across various scientific disciplines. In such settings, an experimental unit may interact with others, leading to potential contamination effects, referred to in this study as network adjustments, which influence the responses of neighboring units. This paper addresses the design problem for connected experimental units subjected to unstructured treatments under linear models, explicitly incorporating network adjustments to account for correlated responses. We employ alphabetic optimality criteria to identify efficient designs that enhance the precision of treatment effect estimation and the accuracy of quantifying network adjustments. Theoretical conditions and practical guidelines for optimal designs are developed and validated through numerical simulations and application to a real-world network. Our findings demonstrate that the proposed approach delivers highly efficient designs while maintaining low computational complexity.

Keywords: Network adjustment; Social network; Mixed-effect; model Optimality

Optimal Experimental Designs for Low-Rank Function Completion

Ming-Hung (Jason) Kao

School of Mathematical & Statistical Sciences, Arizona State University

ABSTRACT

This work is concerned with optimal experimental designs for collecting high-quality sparse functional data, consisting of observations of one or more random functions $X_1(t), \ldots, X_M(t)$ taken at sparse and possibly irregularly spaced points over a compact domain T. Common objectives in analyzing such data include recovering the functions $X_m(t)$ across T, and predicting a response function $Y(t)$ using $X_1(t), \ldots, X_M(t)$ as predictors. Drawing on ideas from low-rank matrix completion, a low-rank function completion (LRFC) framework has recently been proposed to efficiently carry out such analyses. However, optimal design strategies to ensure precise inference under the LRFC framework have not been studied, and existing design approaches for sparse functional data analysis (FDA) may become unwieldy under this framework. In this work, we propose an efficient method for identifying optimal designs tailored to the LRFC framework. We demonstrate its effectiveness and highlight its applicability to studies where other sparse FDA methods are considered.

Keywords: Design efficiency; Sparse functional data analysis; Low-rank approximation; Matrix completion; Optimization algorithms

Efficient Bayesian Estimation and Inference for Shapley Value via Optimal Design

Zheng Zhou[1], Yongdao Zhou[2], Robert Mee[3], **Wei Zheng[3]**

[1]*Department of Statistics, Beijing University of Technology*

[2]*School of Statistics and Data Science, Nankai University*

[3]*Department of Business Analytics and Statistics, University of Tennessee Knoxville*

ABSTRACT

The Shapley value, a fundamental concept in cooperative game theory, provides a fair allocation of cooperative gains or costs among players. However, computing Shapley values for a game with d players requires evaluating all 2^d coalitions, which is computationally infeasible for large d. This difficulty is exacerbated in modern applications such as artificial intelligence, data science, and genomics, where evaluating the value of even a single coalition can be costly. To enable fast approximation and the probabilistic inference of the Shapley value, we propose the Bayesian framework, where the Gaussian process is adopted to infer unobserved coalition values. The posterior distribution of the coalition values are then transformed into that of the Shapley values, allowing both point estimation and uncertainty quantification. We derive theoretical results showing that the computational complexity of posterior evaluation can be reduced from exponential to polynomial order. To further improve efficiency, we integrate experimental design principles to select coalitions that minimize posterior variances. Compared with existing approaches, the proposed method offers three main advantages: (i) accurate estimation of Shapley values using as few as d^2-d+1 coalition evaluations, (ii) robustness to the form of the cooperative game without requiring specific assumptions, and (iii) support for statistical inference in addition to point estimation. Simulation studies and case analyses demonstrate that the proposed approach consistently achieves higher accuracy and efficiency than existing methods under comparable evaluation cost.

Keywords: Bayesian analysis; Data valuation; Design of experiments; Feature selection; Game theory

Cyclic SOAs and Moving Window Criteria for Space-Filling Designs

Wei-Chun, Kang, **Cheng-Yu, Sun**

Institute of Statistics and Data Science, National Tsing Hua University

ABSTRACT

Space-filling designs are essential in computer experiments to ensure that design points are well distributed across the input space. Among them, strong orthogonal arrays (SOAs) are widely used for their stratification properties on fixed grids formed by collapsing adjacent factor levels. However, these grids lack flexibility and cannot adapt to local structures. To address this limitation, we introduce a moving window approach that evaluates uniformity over sliding regions across the space. By specifying a window size and selecting a subset of positions, we define a more fine-grained space-filling criterion that generalizes several existing methods. Within this framework, we further propose cyclic SOAs— SOAs that preserve their stratification properties under cyclic shifts of factor levels. These designs exhibit a structural invariance that is particularly useful in settings involving periodicity or level relabeling. We establish their optimality properties and present construction methods, positioning cyclic SOAs as a flexible and robust addition to the space-filling design toolkit.

Keywords: Computer experiments; Centered L_2-discrepancy; Wrap-around L_2-discrepancy; Minimum Aberration; Cyclic design

Two-Stage Adaptive Testing of Large-Scale Mediation Hypotheses

Yueqi Xu, **Kwun Chuen Gary Chan**

Department of Biostatistics, University of Washington

ABSTRACT

In testing many mediation hypotheses, we propose to use a pair of asymptotically independent p-values constructed from the exposure-mediator and mediator-outcome path-specific p-values from each hypothesis, to test a more stringent joint null hypothesis and a composite mediation null hypothesis respectively. A first-stage screening procedure based on the p-values testing a joint hypothesis can effectively reduce the number of mediation hypotheses to be tested in the second stage. The procedure controls false discovery rate while being consistently well powered across a wide range of scenarios.

Keywords: Composite null hypothesis, False-discovery rate control, Joint significance, Ordered statistics.

Leveraging Multi-Study, Multi-Outcome Data to Improve External Validity and Efficiency of Clinical Trials for Medications for Opioid Use Disorder

Amy Pitts[1], Oliver Hines[2], Rachael Ross[3], Kara Rudolph[2], **Caleb Miles[2]**

Regeneron Pharmaceuticals

Columbia University

University of Utah

ABSTRACT

As data sources have become more plentiful and readily accessible, the practice of data fusion has become increasingly ubiquitous. However, when the focus is on a causal effect on a particular outcome, a major limitation is that this outcome may not be available in all data sources. In fact, different randomized experiments or observational studies of a common exposure will often focus on potentially related yet distinct outcomes. One such example is the medication for opioid use disorder (MOUD) clinical trials network (CTN), which consists of several randomized trials of the comparative effectiveness of different MOUDs with inconsistent quality of life measures across studies. The causally principled methodology is developed for fusing data sets when multiple outcomes are observed across studies, which leverages mediators and outcomes of secondary interest as informative proxies for the missing outcome of primary interest, thereby maximizing power and efficiency by making full use of the available data. As this methodology relies on a key transportability assumption, methods are also developed to assess the degree of sensitivity to violations of this assumption. This methodology is applied to data from the CTN trials to make improved causal inferences about the comparative effectiveness of medications for opioid use disorder.

Keywords: Causal inference, data fusion, external validity, generalizability, missing data, transportability

Mediation Analysis with Graph Mediator

Yixi Xu[1], **Yi Zhao[1,*]**

[1]*Department of Biostatistics and Health Data Science, Indiana University School of Medicine*

ABSTRACT

This study introduces a mediation analysis framework when the mediator is a graph. A Gaussian covariance graph model is assumed for graph presentation. Causal estimands and assumptions are discussed under this presentation. With a covariance matrix as the mediator, a low-rank representation is introduced and parametric mediation models are considered under the structural equation modeling framework. Assuming Gaussian random errors, likelihood-based estimators are introduced to simultaneously identify the low-rank representation and causal parameters. An efficient computational algorithm is proposed and asymptotic properties of the estimators are investigated. Via simulation studies, the performance of the proposed approach is evaluated. Applying to a resting-state fMRI study, a brain network is identified within which functional connectivity mediates the sex difference in the performance of a motor task.

Keywords: Common diagonalization; Covariance regression; Decomposition method; Gaussian covariance graph model; Mediation analysis

Toward Flexible and Efficient Counterfactual Density Estimation

Kwangho Kim[1], Jisu Kim[2], Edward Kennedy[3]

[1]*Department of Statistics, Korea University*

[2]*Department of Statistics, Seoul National University*

[3]*Department of Statistics and Data Science, Carnegie Mellon University*

ABSTRACT

Comparing full counterfactual distributions provides richer measures of causal effects than conventional summaries such as means or quantiles. We study the problem of estimating the entire counterfactual density in a fully nonparametric setting and develop three complementary approaches. First, we introduce a doubly robust–style estimator that directly targets a kernel–smoothed counterfactual density. We establish its large-sample properties, derive finite-sample risk bounds, and construct uniform confidence bands via a bootstrap procedure. Second, we propose a diffusion-informed bump that adapts to the intrinsic geometry of the outcome manifold. By replacing the standard kernel with a diffusion-informed smoother, this estimator reduces bias near complex supports and attains faster convergence rates for high-dimensional outcomes when intrinsic dimension is low. Third, we develop a score-based method that targets the smoothed counterfactual score rather than the density itself. Focusing on the score enables even faster rates under common structural assumptions, with smaller constant factors, and can be paired with efficient density recovery when desired. We compare the three estimators and clarify when each is preferable, with particular emphasis on the advantages of diffusion-based smoothing for learning counterfactual distributions in high-dimensional settings. Together, these results provide a unified toolkit for flexible and statistically efficient counterfactual density estimation.

Keywords: causal inference; density estimation; influence function; semiparametric theory; diffusion model

Mengyi Lu

Xin Chen

Learning and Inference for Low-Rank Models

Dong Xia[1]

Department of Mathematics, Hong Kong University of Science and Technology

ABSTRACT

My talk will focus on the learning and inference for low-rank matrices and tensors in the presence of missing values, heterogeneity, and adaptively collected data. These problems pose significant challenges because the estimators are typically derived using iterative algorithms and involve multiple stages of spectral decomposition. Over the past several years, we have made several contributions in this field, including specially crafted estimation procedures, a powerful spectral representation formula, the double-sample debiasing approach, false-discovery control in multiple hypothesis testing, and debiasing using inverse propensity weighting.

Keywords: Low-rank, tensor, inference, spectral methods

Federated Community Detection in Bipartite Networks from Various Platforms

Wanjie Wang

Department of Statistics and Data Science, National University of Singapore

ABSTRACT

In modern applications such as recommendation systems and e-commerce platforms, user-item interactions are naturally modelled by bipartite networks. When data are collected across multiple servers, privacy concerns and communication constraints render centralized community detection impractical. It motivated the development of federated learning. In this talk, I will present a federated framework for community detection in bipartite networks, where user data reside on disjoint servers, but all servers share a common item set.

We introduce a gradient-based power iteration method that approximates leading singular vectors using only local computations and limited inter-server communication. Building on this, we propose the BiFLICKER algorithm, which combines spectral methods with federated learning to recover global community structures while preserving data locality. Theoretical guarantees are established for the accuracy of spectral estimation and community recovery. We validate our method through simulations and real-world analysis of federated movie rating data from Douban, revealing meaningful and interpretable user and movie communities.

Keywords: Federated Learning; Spectral Analysis; Bipartite Network; Community Detection

Online Tensor Inference

Xin Wen, Will Wei Sun, **Yichen Zhang**

Department of Quantitative Methods, Daniels School of Business, Purdue University

ABSTRACT

Contemporary applications, such as recommendation systems and mobile health monitoring, require real-time processing and analysis of sequentially arriving high-dimensional tensor data. Traditional offline learning, involving the storage and utilization of all data in each computational iteration, becomes impractical for these tasks. Furthermore, existing low-rank tensor methods lack the capability for online statistical inference, which is essential for real-time predictions and informed decision-making. This paper addresses these challenges by introducing a novel online inference framework for low-rank tensors. Our approach employs Stochastic Gradient Descent (SGD) to enable efficient real-time data processing without extensive memory requirements. We establish a non-asymptotic convergence result for the online low-rank SGD estimator, nearly matches the minimax optimal estimation error rate of offline models. Furthermore, we propose a simple yet powerful online debiasing approach for sequential statistical inference. The entire online procedure, covering both estimation and inference, eliminates the need for data splitting or storing historical data, making it suitable for on-the-fly hypothesis testing. In our analysis, we control the sum of constructed super-martingales to ensure estimates along the entire solution path remain within the benign region. Additionally, a novel spectral representation tool is employed to address statistical dependencies among iterative estimates, establishing the desired asymptotic normality.

Keywords: Low-rank tensors; online learning; statistical inference; stochastic gradient descent

Generalized Tensor Completion with Non-Random Missingness

Maoyu Zhang[1*], Biao Cai[2*], Will Wei Sun[3], **Jingfei Zhang[1]**

[1]*Goizueta Business School, Emory University, Atlanta, GA, USA.*

[2]*Department of Decision Analytics and Operations, City University of Hong Kong, Hong Kong.*

[3]*Daniels School of Business, Purdue University, West Lafayette, IN, USA.*

Abstract

Tensor completion plays a crucial role in applications such as recommender systems and medical imaging, where data are often highly incomplete. While extensive prior work has addressed tensor completion with data missingness, most assume that each entry of the tensor is available independently with probability p. However, real-world tensor data often exhibit missing-notat-random (MNAR) patterns, where the probability of missingness depends on the underlying tensor values. This paper introduces a generalized tensor completion framework for noisy data with MNAR, where the observation probability is modeled as a function of underlying tensor values. Our flexible framework accommodates various tensor data types, such as continuous, binary and count data. For model estimation, we develop an alternating maximization algorithm and derive non-asymptotic error bounds for the estimator at each iteration, under considerably relaxed conditions on the observation probabilities. Additionally, we propose a statistical inference procedure to test whether observation probabilities depend on underlying tensor values, offering a formal assessment of the missingness assumption within our modeling framework. The utility and efficacy of our approach are demonstrated through comparative simulation studies and analyses of two real-world datasets.

Keywords: generalized tensor completion; low-rank tensor model; missing not at random; non-convex optimization; hypothesis testing.

Diffusion Models for High-Dimensional Digital Twins

Georg Gottwald, Shuigen Liu, Yang Lyu, Youssef Marzouk, Tan Nguyen, Yuchun Qian, Sebastian Reich, **Xin T. Tong**

Department of Mathematics, National University of Singapore

ABSTRACT

Diffusion model is a popular tool to generate new data samples with possible applications in digital twin training. However, rigorous understanding of the diffusion model is still lacking. One issue is how to train these models for high dimensional problems as score function estimation is subject to the curse of dimension. Another issue is how to avoid the memorization effect, where the diffusion model is bound to generate an exact copy from the training data. We will provide solutions to the first issue by focusing on high dimensional distributions with sparse dependence. We will leverage the sparse dependence to provide a local estimation of the score functions. As for the second issue, we will modify the diffusion model in the final stage and generate new samples close to the same manifold where the training data is originated.

Keywords: Diffusion models; Curse of dimension; manifold hypothesis, Sparse dependence

Two-Sample Tests for Equal Distributions in Separable Metric Spaces: A Unified Semimetric-Based Approach

Jin-Ting Zhang[1], Meichen Qian[1], and Tianming Zhu[2]

[1]*Department of Statistics and Data Science, National University of Singapore, Singapore*

[2]*National Institute of Education, Nanyang Technological University, Singapore*

ABSTRACT

With the advancement of data collection techniques, researchers frequently encounter complex data objects within separable metric spaces across various domains. One common interest lies in determining whether two groups of complex data objects originate from the same population. This paper introduces and examines a fast and accurate unified semimetric-based approach designed to tackle this challenge. The approach exhibits broad applicability across a wide range of research areas, such as bioinformatics, audiology, environmentology, finance, and more. It effectively identifies differences between the distributions of two complex datasets, including both high-dimensional data and functional data. The asymptotic null and alternative distributions of the proposed test statistic are established. Unlike the permutation approach, a unified, rapid and precise method to approximate the null distribution is described. Furthermore, the proposed test is shown to be root-n consistent. Numerical results are presented for illustrating the excellent performance of the proposed test in terms of size control, power, and computational cost. Additionally, the applications of the proposed test are showcased through examples involving both high-dimensional data and functional data.

Keywords: Two-sample test; Equal distribution; Unified semimetric-based approach; Three-cumulant matched chi-square-approximation

A Fast and Accurate Kernel-Based Independence Test with Applications to High-Dimensional and Functional Data

Jin-Ting Zhang[1], **Tianming Zhu[2,*]**

[1]*Department of Statistics and Data Science, National University of Singapore*

[2]*National Institute of Education, Nanyang Technological University*

ABSTRACT

Testing the dependency between two random variables is an important inference problem in statistics since many statistical procedures rely on the assumption that the two samples are independent. To test whether two samples are independent, a so-called HSIC (Hilbert--Schmidt Independence Criterion)-based test has been proposed. Its null distribution is approximated either by permutation or a Gamma approximation. In this paper, a new HSIC-based test is proposed. Its asymptotic null and alternative distributions are established. It is shown that the proposed test is root-n consistent. A three-cumulant matched chi-squared-approximation is adopted to approximate the null distribution of the test statistic. By choosing a proper reproducing kernel, the proposed test can be applied to many different types of data including multivariate, high-dimensional, and functional data. Three simulation studies and two real data applications show that in terms of level accuracy, power, and computational cost, the proposed test outperforms several existing tests for multivariate, high-dimensional, and functional data.

Keywords: complicated data objects; Hilbert-Schmidt independence criterion; three-cumulant matched chi-squared-approximation; two-sample independence test

Bayesian Extrapolation Design: Exposure-Response Curve Comparison between Pediatric and Adult Populations

Jingjing Ye

BeOne Medicines

ABSTRACT

Developing effective treatments for pediatric populations presents unique scientific and ethical challenges, particularly given the small population size. Both U.S. and EU regulations advocate for pediatric extrapolation, a strategy that leverages existing adult data to assess its relevance to children. This approach often depends on demonstrating similar disease progression, pharmacology, and clinical responses between adults and children. In pharmacology, similarity is typically evaluated through the exposure-response (E-R) relationship. However, current methods for comparing E-R curves between these groups are limited, often focusing on isolated data points rather than the entire curve (Zhang et al., 2021).

To address this gap, we propose an innovative Bayesian approach for a comprehensive comparison of E-R curves between adults and pediatric populations. This method evaluates the full spectrum of the curve using logistic regression for binary endpoints. We developed an algorithm to determine optimal sample size and key design parameters, including the Bayesian posterior probability threshold, and use the maximum curve distance as a similarity metric.
By integrating Bayesian and frequentist principles, our approach simulates datasets under both null and alternative hypotheses, ensuring type I error control while maximizing statistical power. Simulation studies demonstrate that this method offers better type I error control and greater power compared to traditional frequentist approaches (Dette et al., 2018).

Keywords: Bayesian Design; Extrapolation; Exposure-Response; Similarity Assessment.

Bayesian Inference for Cluster-Randomized Trials with Multivariate Outcomes Subject to Both Truncation by Death and Missingness

Guangyu Tong[1234], Chenxi Li[5], Eric Velazquez[1], Michael O. Harhay[67], Fan Li[124]

[1]*Department of Internal Medicine, Section of Cardiovascular Medicine, Yale School of Medicine, New Haven, CT, USA*

[2]*Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, USA*

[3]*Cardiovascular Medicine Analytics Center, Yale School of Medicine, New Haven, Connecticut, USA*

[4]*Center for Methods in Implementation and Prevention Science, Yale University, New Haven, Connecticut, USA*

[5]*Department of Biostatistics & Bioinformatics, Duke School of Medicine, Durham, North Carolina, USA*

[6]*Clinical Trials Methods and Outcomes Lab, Palliative and Advanced Illness Research (PAIR) Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA*

[7]*Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA*

ABSTRACT

Cluster-randomized trials (CRTs) on fragile populations often face complex attrition, where missing outcomes arise from varied causes—participants may be known alive, known deceased, or have unknown survival status. Existing CRT methods handle truncation by death but do not accommodate participants who drop out for unrelated reasons or with unascertained survival. We propose a Bayesian framework for estimating survivor average causal effects (SACE) in CRTs that accounts for different types of missingness. Our approach jointly estimates causal effects using multivariate outcomes and distinguishes between individual-level and cluster-level SACE. Through simulation studies, we show low bias and high coverage across scenarios. We also apply the method to a geriatric CRT, illustrating its use with bivariate continuous outcomes. The framework extends naturally to multiple or non-continuous endpoints, providing a general solution for handling complex missingness in CRTs, especially in aging and palliative care populations.

Keywords: Bayesian inference; survivor average causal effect; survivor cluster-average causal effect; informative cluster size

Design of a Dual Randomized Trial in a Type 2 Hybrid Effectiveness-Implementation Study

Feng-Chang Lin

Department of Biostatistics, University of North Carolina at Chapel Hill

ABSTRACT

Dual randomized controlled trials (DRCT) are type 2 hybrid studies that include two randomized trials: one testing implementation strategies and one testing an intervention. We argue that this study design offers efficiency by providing rigorous investigation of both implementation and intervention in one study and has potential to accelerate generation of the evidence needed to translate interventions that work into real-world practice. Nevertheless, studies using this design are rare in literature. We construct a paradigm that breaks down the components of the DRCT and provide a step-by-step explanation of features of the design and recommendations for use. A clear distinction is made between the dual strands that test the implementation versus the intervention, and a minimum of three randomized arms is advocated. We suggest an active treatment arm that includes both the implementation strategy and intervention that are hypothesized to be superior. We suggest two comparison/control arms: one to test the implementation strategy and the second to test the intervention. Further, we recommend selection criteria for the two control arms that place emphasis on maximizing the utility of the study design to advance public health practice. On the surface, the design of a DRCT can appear simple, but actual application is complex. We believe it is that complexity that has limited its use in literature. We hope that this paper will give both implementation scientists and trialists who are not familiar with implementation science a better understanding of the DRCT design and encouragement to use it.

Keywords: Randomized trial, Hybrid effectiveness-implementation study, Implementation science

ProjectDRIVE: Effect of In-Vehicle Feedback with and without Parent Communication Training on Teen Driving Behaviours

Jingzhen Yang, PhD, MPH[1,2], Ying Zhang, PhD[3], Enas Alshaikh, PhD[1], Hannah Schneider, MPH[1], Archana Kaur, MPH[1], Dominique M. Rose, PhD, MPH[1], Priyanka Sridharan, MPH[1], Armita Kar, PhD[6], Kele Ding, PhD, MD[1,4], Yang Wang, PhD[5], Xueyuan Ren, MS[5], Miao Yu, MS[5], Lisa Roth, MPH[7], Cara Hamann, PhD[7,8], Elizabeth E. O'Neal, PhD[7,9], Motao Zhu, PhD, MD[1,2], Corinne Peek-Asa, PhD, MPH[10]

[1]Center for Injury Research and Policy at the Abigail Wexner Research Institute, Nationwide Children's Hospital, Columbus, Ohio, USA

[2]Department of Pediatrics, The Ohio State University, Columbus, Ohio, USA

[3]Department of Biostatistics, University of Nebraska Medical Center, Omaha, Nebraska, USA

[4]Health Education and Promotion Program, School of Health Sciences, Kent State University, Kent, Ohio, USA

[5]Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, USA

[6]Department of Geography and Geoinformation Science, George Mason University, Fairfax, Virginia, USA

[7]University of Iowa Injury Prevention Research Center, Iowa City, Iowa, USA

[8]Department of Epidemiology, University of Iowa, Iowa City, Iowa, USA

[9]Department of Community and Behavioral Health, University of Iowa, Iowa City, Iowa, USA

[10]Office of Research Affairs, University of California at San Diego, San Diego, California, USA

ABSTRACT

Motor vehicle collisions are a leading cause of death among U.S. teens, and those with traffic violations are at particularly high risk. Despite this, few evidence-based interventions target unsafe driving in this population. Parent involvement shows promise but is rarely incorporated into post-licensure safety programs. This three-arm, parallel-group randomized controlled trial evaluated the effectiveness of *ProjectDRIVE*, an in-vehicle feedback intervention delivered with and without parent communication training, in reducing risky driving among teen drivers with traffic violations. We enrolled 240 parent–teen dyads from six juvenile traffic courts across

Ohio (September 2020–June 2024). Teens were 16–17 years old, held an intermediate driver's license, and had a recent moving violation. Dyads were randomized equally (n=80 per arm) to: 1) **Control**: telematics device installed without feedback; 2) **Driving Feedback Only**: teens received real-time in-vehicle alerts, an app-based dashboard, and bi-weekly emailed driving reports; or 3) **Driving Feedback + Parent Training**: teens received the same feedback as arm 2, and parents accessed driving reports and completed a virtual communication training module. Intervention exposure lasted 6 months. Telematics captured risky driving events—hard braking, sudden acceleration, speeding >10 mph over the limit, and speeds >75 mph. The primary outcome was the incidence rate of risky events per 1,000 miles driven. Teens completed 160,095 trips. Driving Feedback alone did not significantly reduce risky driving. In contrast, Driving Feedback + Parent Training significantly lowered overall risky driving event rates compared with controls (aIRR 0.68; 97.5% CI, 0.51–0.90). Reductions were also observed for hard braking (aIRR 0.77) and speeding >75 mph (aIRR 0.81). Male teens consistently exhibited higher risky driving rates. In conclusion, combining driving feedback with structured parent training effectively reduced risky driving among teens with traffic violations. Sustained parental engagement may be critical for improving post-licensure teen driving safety.

Keywords: Teen drivers, In-vehicle technology, Parental communication, Telematics, Randomized controlled trial

Back to Sessions List

Disease Progression Modeling with Informative Censoring

Thomas Jensen, MS

Berry Consultants, Utah State University

ABSTRACT

Clinical studies in progressive diseases encounter missing outcome data when trial participants exit the trial before completing the full specified follow up duration. Unlike time to event analyses which naturally handle missingness through censoring, repeated measures analyses of disease progression may be subject to bias when missing data are improperly handled. Commonly used mixed models for repeated measures generally assume missingness is random, or in other words, unrelated to the study outcome. However, in many progressive diseases, such as interstitial lung disease (ILD), missing data are often highly correlated with bad outcomes. For instance, ILD trial participants who are in late stages of disease progression manifest by poor performance on the outcome scale are generally understood to be more likely to have missing data due to a mortality event than healthy participants. While most analysis plans will pre-specify secondary analyses of informative event times, conducting separate analyses may increase the possibility of observing conflicting trial results. The ideal repeated measures analysis directly accounts for potentially informative missingness in modeling disease progression.

Joint models of repeated measure and time to event outcomes allow for proper handling of informative missingness within the mixed model framework. In this talk, the general structure of joint models will be explored, coupled with examples highlighting plausible sources of bias in mixed model analyses and corresponding bias correction with joint model analyses. A simulation study loosely based on publicly available ILD trial data will be presented and results including expected bias and statistical power will be discussed.

Keywords: disease progression modeling, informative censoring, joint models

Change Surface Regression for Nonlinear Subgroup Identification

Jialiang Li

Department of Statistics and Data Science, National University of Singapore

ABSTRACT

Pharmacogenomics stands as a pivotal driver toward personalized medicine, aiming to optimize drug efficacy while minimizing adverse effects by uncovering the impact of genetic variations on inter-individual outcome variability. Despite its promise, the intricate landscape of drug metabolism introduces complexity, where the correlation between drug response and genes can be shaped by numerous non-genetic factors, often exhibiting heterogeneity across diverse subpopulations. This challenge is particularly pronounced in datasets like the International Warfarin Pharmacogenetic Consortium (IWPC), which encompasses diverse patient information from multiple nations. To capture the between-patient heterogeneity in dosing requirement, we formulate a novel change surface model as a model-based approach for multiple subgroup identification in complex datasets. A key feature of our approach is its ability to accommodate nonlinear subgroup divisions, providing a clearer understanding of dynamic drug-gene associations. Furthermore, our model effectively handles high-dimensional data through a doubly penalized approach, ensuring both interpretability and adaptability. We propose an iterative two-stage method that combines a change point detection technique in the first stage with a smoothed local adaptive majorize-minimization algorithm for surface regression in the second stage. Performance of the proposed methods is evaluated through extensive numerical studies.

Application of our method to the IWPC dataset leads to significant new findings, where three subgroups subject to different pharmacogenomic relationships are identified, contributing valuable insights into the complex dynamics of drug-gene associations in patients.

Keywords: Personalized medicine; subgroup identification; change point detection

Unraveling the Neurogenetic Architecture of Nicotine Use and Depression: A Network-Driven Integration of Brain Transcriptomes and GWAS

Bao-Zhu Yang[1], Bo Xiang[2], Yidong Li[1], Daniel Levey[1,3], Uri Bright[1,3], Tingting Wang[2], Chiang-Shan Li[1,4,5], Marc N. Potenza[1,4,5,6,7,8], Shuangge Ma[9], Joel Gelernter[1,3,4,5,10]

[1]*Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA*

[2]*Department of Psychiatry, Fundamental and Clinical Research on Mental Disorders Key Laboratory of Luzhou, Laboratory of Neurological Diseases & Brain Function, Affiliated Hospital of Southwest Medical University, Luzhou, Sichuan Province, China*

[3]*Veterans Affairs Connecticut Healthcare Center, West Haven, CT, USA*

[4]*Department of Neuroscience, Yale University, New Haven, CT, USA*

[5]*Wu Tsai Institute, Yale University School of Medicine, New Haven, CT, USA*

[6]*Child Study Center, Yale School of Medicine, New Haven, CT, USA*

[7]*Connecticut Council on Problem Gambling, Wethersfield, CT*

[8]*Connecticut Mental Health Center, New Haven, CT*

[9]*Department of Biostatistics, Yale University, New Haven, CT, USA*

[10]*Department of Genetics, Yale University School of Medicine, New Haven, CT, USA*

ABSTRACT

This study employed a network-based framework to investigate the neurogenetic mechanisms underlying individual variation in complex traits. We constructed whole-brain and intramodular region-specific coexpression networks using BrainSpan transcriptomes, integrated genetic risk factors through genome-wide association studies, validated the associations between genetic risk and each gene set in the region-specific subnetworks using an independent dataset, and conducted bioinformatic analyses. To illustrate this approach, we applied it to two case studies: (1) nicotine use severity, where we examined the genetic risks associated with cigarettes per day and their relationship to brain coexpression networks, and (2) major depressive disorder, where we explored transcriptomic and genetic contributions to depression-related neurobiological pathways. These applications demonstrate the utility of our network-based approach in uncovering neurogenetic mechanisms underlying psychiatric and substance use

disorders.

The cigarettes per day project was published in 2024.
https://pubmed.ncbi.nlm.nih.gov/38422867/

The major depressive disorder project was presented as a poster in October 2024 at the World Congress of Psychiatric Genetics, held in Singapore. The corresponding manuscript is currently under review.

Keywords: Neurogenetic architecture; Network analysis; Brain transcriptome; GWAS

Back to Sessions List

Enhancing Model Generalizability in Medical AI: Systematic Comparison of Categorical Encoding and Sampling Techniques for Imbalanced Data

Chien-Wei Chuang[1,†], Chung-Kuan Wu[2,3,4,†], Ben-Chang Shia[1], Mingchih Chen[1,*]

[1]*Graduate Institute of Business Administration, Fu Jen Catholic University, New Taipei City, Taiwan*

[2]*Division of Nephrology, Shin Kong Wu Ho-Su Memorial Hospital, Taipei, Taiwan*

[3]*Dialysis Access Management Center, Shin Kong Wu Ho-Su Memorial Hospital, Taipei, Taiwan*

[4]*School of Medicine, Fu Jen Catholic University, New Taipei, Taiwan*

**Correspondence: 081438@mail.fju.edu.tw*

[†]*Co-first author, these authors contributed equally to this work*

ABSTRACT

Background: Despite the increasing use of machine learning (ML) in clinical research, the early stages of data preparation especially for structured clinical data often receive limited methodological scrutiny. These datasets typically contain missing values, complex categorical variables, and imbalanced class distributions, all of which complicate downstream model development and interpretation.

Objective: This study introduces a structured preprocessing framework designed to address common challenges in medical tabular data and to assess how preprocessing choices affect the stability and portability of predictive models across settings.

Methods: We constructed a modular workflow comprising three components. First, preprocessing strategies included imputation for missing data, three types of categorical encoding (One-Hot, Frequency, Target), and resampling approaches for class imbalance (SMOTE, ROSE). Second, six classification algorithms were used to evaluate performance patterns: Logistic Regression, Decision Tree, Random Forest, XGBoost, CatBoost, and LightGBM. Third, we assessed cross-dataset portability using two datasets with distinct data-generating mechanisms: an ESRD patient registry (n=412) and the population-based BRFSS 2015 survey. For each dataset, we independently cleaned, standardized, encoded, tuned, and evaluated models, without cross-dataset feature matching or pooled AUC calculations, and re-ran the complete pipeline on BRFSS as an external replication.

Results: One-Hot Encoding in combination with ROSE yielded the most consistent performance improvements in terms of AUC (0.940) and accuracy (0.932), particularly for

classifiers sensitive to class distribution. Notably, ROSE enhanced sensitivity without substantially distorting the original data structure. Feature importance rankings also contributed to model interpretability, and performance trends were largely reproducible in cross-context application.

Conclusions: Our findings suggest that preprocessing decisions often treated as ancillary play a central role in shaping model outcomes, especially in high-variance clinical datasets. The proposed framework offers a reproducible and adaptable tool for aligning data preparation with the unique demands of healthcare prediction tasks, and may serve as a foundation for future efforts to standardize preprocessing in clinical ML workflows.

Keywords: Machine Learning; Data Preprocessing; Clinical Prediction Models; Medical Informatics; Feature Engineering

Back to Sessions List

Time-Varying Latent Space Modeling for Outcome-Based Human Disease Network

Guojun Zhu[1], Ruiyue Wan[1]g, Rong Li[2], Sanguo Zhang[1], Shuangge Ma[2], Guanzhong Qiao[3], **Hao Mei**[4]

[1]*School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China*

[2]*Department of Biostatistics, Yale School of Public Health, New Haven, USA*

[3]*Department of Orthopaedic, The First Hospital of Tsinghua University, Beijing, China*

[4]*Center for Applied Statistics, School of Statistics, Institute of Health Data Science, Renmin University of China, Beijing, China*

ABSTRACT

Human disease network (HDN) analysis, which jointly considers a large number of diseases and focuses on their interconnections, is getting increasingly popular and can shed important insight not possessed by individual-disease-based analysis. Multiple network analysis techniques have been developed for HDNs, although new developments are still strongly needed.

In this article, we adopt latent space modeling, which has proven powerful in other network analysis contexts and offers unique, insightful interpretations, but has been limitedly applied in HDN analysis. Different from some other types of network analysis and some other HDN analyses (such as gene-centric ones), in this article, we pay unique attention to modeling temporal variations. For this purpose, a penalization approach is developed, which can identify time regions with constant network structures (that correspond to ignorable changes) as well as those with smooth variations. The statistical and computational properties are rigorously established. With Medicare data -- one of the most powerful medical claims databases -- we analyze the admission records of 133 million hospital inpatient treatments from January 2008 to December 2019. Sensible findings are made on disease interconnections and clustering structures. Additionally, the temporal variations, which have not been revealed in the literature, are found to be interpretable. The analysis can provide a new way for connecting and grouping diseases and assist in understanding and planning medical resources.

Keywords: Human disease network; Latent space modeling; Temporal variation; Penalized estimation; Medicare data

AI-Powered Bayesian Methods for Analyzing Spatial Omics Data

Qiwei Li[1]

Department of Mathematical Sciences, The University of Texas at Dallas

ABSTRACT

Recent technology breakthroughs in gene expression profiling have enabled the comprehensive molecular characterization of single cells while preserving their spatial and morphological contexts. This new bioinformatics scenario advances our understanding of molecular and cellular spatial organizations in tissues, fuelling the next generation of scientific discovery. Bayesian statistics relies more on human analyses with computer aids, while AI relies more on computer algorithms with aids from humans. This talk will outline methodologies for bridging AI capabilities with Bayesian frameworks, aiming to resolve key issues in spatially resolved transcriptomics (SRT) data analysis. Particularly, I will focus on two major problems: spatial domain identification and gene expression reconstruction.

Current clustering analysis of SRT data primarily relies on molecular information and fails to fully exploit the morphological features present in histopathology images, leading to compromised accuracy and interpretability. To overcome these limitations, we have developed a multi-stage statistical method called iIMPACT. It identifies and defines histology-based spatial domains based on AI-reconstructed histopathology images and spatial context of gene expression measurements, and detects domain-specific differentially expressed genes. Through multiple case studies, we demonstrate iIMPACT outperforms existing methods in accuracy and interpretability and provides insights into the cellular spatial organization and landscape of functional genes within spatial transcriptomics data.

Most next-generation sequencing-based SRT techniques are limited to measuring gene expression in a confined array of spots, capturing only a fraction of the spatial domain. Typically, these spots encompass gene expression from a few to hundreds of cells, underscoring a critical need for more detailed, single-cell resolution SRT data to enhance our understanding of biological functions within the tissue context. Addressing this challenge, we introduce BayesDeep, a novel Bayesian hierarchical model that leverages cellular morphological data from histology images, commonly paired with SRT data, to reconstruct SRT data at the single-cell resolution. BayesDeep effectively models count data from SRT studies via a negative binomial regression model. This model incorporates explanatory variables such as cell types and nuclei-shape information for each cell extracted from the paired histology image. A feature selection scheme is integrated to examine the association between the morphological and

molecular profiles, thereby improving the model robustness. We applied BayesDeep to two real SRT datasets, successfully demonstrating its capability to reconstruct SRT data at the single-cell resolution. This advancement not only yields new biological insights but also significantly enhances various downstream analyses, such as pseudotime and cell-cell communication.

Keywords: Spatial transcriptomics; Spatial clustering; Shape analysis; Deep learning

Back to Sessions List

Covariate-Assisted Graph Matching

Trisha Dawn, **Jesús Arroyo**

Department of Statistics, Texas A&M University

ABSTRACT

Data integration is essential across diverse domains, from historical records to biomedical research, facilitating joint statistical inference. A crucial initial step in this process involves merging multiple data sources based on matching individual records, often in the absence of unique identifiers. When the datasets are network data -- a flexible and increasingly prevalent structure representing entities as vertices and their relationships as edges -- this problem is typically addressed through graph matching methodologies. For such cases, auxiliary features or covariates associated with nodes or edges can be instrumental in achieving improved accuracy. However, most existing graph matching techniques do not incorporate this information, limiting their performance against non-identifiable and erroneous matches. To overcome these limitations, we propose two novel covariate-assisted seeded graph matching methods, where a partial alignment for a set of nodes, called seeds, is known. The first one utilizes the quadratic assignment problem (QAP), while the second one leverages the local neighborhood structure of non-seed nodes to guide the matching process. Both methods are grounded in a conditional modeling framework, where elements of one graph's adjacency matrix are modeled using a generalized linear model (GLM), given the other graph and the available covariates. We establish theoretical guarantees for model estimation error and exact recovery of the solution of the QAP, demonstrating perfect alignment accuracy with high probability under sufficient signal strength. The effectiveness of our methods is demonstrated through numerical experiments to accommodate varied parameter settings and number of seeds. Finally, we apply our proposed approach to match two real-world network data. Our work highlights the power of integrating covariate information in the classical graph matching setup, offering a practical and improved framework for combining network data with wide-ranging applications.

Keywords: Network data; node and edge covariates; optimization problem; generalized linear model

Abhishek Roy

Seeing Is Believing: Challenges and Opportunities for Super-Resolution Microscopy Image Data Analysis for Quantitative Molecular Biology

Chongzhi Zang
University of Virginia

ABSTRACT

Advanced technologies have been the driving force behind modern biomedical research, with rigorous data analysis serving as the critical link between technological applications and scientific discoveries. While high-throughput sequencing-based technologies have advanced to the single-cell level with spatial information, they still lack the ability to directly measure the dynamic structures and activities of protein molecules in the cell nucleus. Super-resolution microscopy technologies, such as three-dimensional structured illumination microscopy (3D-SIM) and 3D stochastic optical reconstruction microscopy (STORM), have been increasingly applied in molecular biology research, as they enable the direct detection of 3D structures of chromatin elements at the single-molecule resolution. However, data analysis remains a major challenge in this field due to the noisy, sparse, and highly variable nature of microscopy image data across samples, whether in replicates or under different biological conditions. In this talk, I will introduce super-resolution microscopy technologies, discuss the unique characteristics of the big data and challenges of the analysis, and emphasize the need for innovative and rigorous statistical methods to better understand these image data to elucidate molecular biology mechanisms.

Keywords: Image analysis, super-resolution microscopy, bioinformatics

Optimal Score Estimation via Empirical Bayes Smoothing

Andre Wibisono, **Yihong Wu**, Kaylee Yingxi Yang

Yale University

ABSTRACT

We study the problem of estimating the score function of an unknown probability distribution ρ^* from n independent and identically distributed observations in d dimensions. Assuming that ρ^* is subgaussian and has a Lipschitz-continuous score function s^*, we establish the optimal rate of $\widetilde{\Theta}(n^{-\frac{2}{d+4}})$ for this estimation problem under the loss function $\| \hat{s} - s^* \|^2_{L2(\rho^*)}$ that is commonly used in the score matching literature, highlighting the curse of dimensionality where sample complexity for accurate score estimation grows exponentially with the dimension d. Leveraging key insights in empirical Bayes theory as well as a new convergence rate of smoothed empirical distribution in Hellinger distance, we show that a regularized score estimator based on a Gaussian kernel attains this rate, shown optimal by a matching minimax lower bound. We also discuss extensions to estimating β-Hölder continuous scores with $\beta \leq 1$, as well as the implication of our theory on the sample complexity of score-based generative models.

Keywords: Score estimation, kernel density estimation, empirical Bayes, Hellinger distance

Provable Statistical and Computational Efficiency of Diffusion Models

Changxiao Cai, Gen Li

University of Michigan, Chinese University of Hong Kong

ABSTRACT

Score-based diffusion models have emerged as a foundational paradigm for modern generative modeling, achieving remarkable success across diverse applications from image synthesis to scientific computing. Despite their empirical prominence, fundamental questions about their theoretical foundations remain: How efficient can diffusion samplers be? What are the fundamental statistical limits of these samplers? In this talk, I will present recent theoretical advances that address both the computational and statistical frontiers of diffusion models. First, I will introduce a novel accelerated stochastic sampler that provably reduces iteration complexity under minimal assumptions, offering sampling speedups without sacrificing statistical optimality. Second, I will present the first comprehensive end-to-end analysis for deterministic ODE-based samplers, establishing (near-)minimax optimal statistical guarantees under mild assumptions on the target distribution. Together, these results provide a rigorous mathematical foundation that narrows the gap between the practical success and theoretical understanding of diffusion models. This is joint work with Gen Li.

Paper 1: https://arxiv.org/abs/2410.23285

Paper 2: https://arxiv.org/abs/2503.09583

Keywords: diffusion models, sampling, minimax optimiality, probability flow ODE, training-free acceleration

Transformers Provably Learn Chain-of-Thought Reasoning with Length Generalization

Yuejie Chi

Department of Statistics and Data Science, Yale University

ABSTRACT

The ability to reason lies at the core of artificial intelligence (AI), and challenging problems usually call for deeper and longer reasoning to tackle. A crucial question about AI reasoning is whether models can extrapolate learned reasoning patterns to solve harder tasks with longer chain-of-thought (CoT). In this work, we present a theoretical analysis of transformers learning on synthetic state-tracking tasks with gradient descent. We mathematically prove how the algebraic structure of state-tracking problems governs the degree of extrapolation of the learned CoT. Specifically, our theory characterizes the length generalization of transformers through the mechanism of attention concentration, linking the retrieval robustness of the attention layer to the state-tracking task structure of long-context reasoning. Moreover, for transformers with limited reasoning length, we prove that a recursive self-training scheme can progressively extend the range of solvable problem lengths. To our knowledge, we provide the first optimization guarantee that constant-depth transformers provably learn -complete problems with CoT, significantly going beyond prior art confined in , unless the widely held conjecture fails. Finally, we present a broad set of experiments supporting our theoretical results, confirming the length generalization behaviors and the mechanism of attention concentration.

Keywords: transformers, chain-of-thought, length generalization

Mean Shift for Clustering Functional Data: A Scalable Algorithm and Convergence Analysis

Ting-Li Chen[1], **Toshinari Morimoto[2]**, Su-Yun Huang[1], Ruey S. Tsay[3]

[1]*Institute of Statistical Science, Academia Sinica*

[2]*Department of Mathematics, National Taiwan University*

[3]*Booth School of Business, University of Chicago*

ABSTRACT

This paper extends the mean shift algorithm from vector-valued data to functional data, enabling effective clustering in infinite-dimensional settings. To address the computational challenges posed by large-scale datasets, we introduce a fast stochastic variant that significantly reduces computational complexity. We provide a rigorous analysis of convergence for the full functional mean shift procedure, establishing theoretical guarantees for its behavior. For the stochastic variant, we provide some partial justification for its use by showing that it approximates the full algorithm well when the subset size is sufficiently large. The proposed method is validated both through simulation studies and through real-data analysis, including hourly Taiwan PM2.5 measurements and Argo oceanographic profiles. Our key contributions include:

(1) a novel extension of the mean shift algorithm to functional data for clustering without the need to specify the number of clusters

(2) convergence analysis of the full functional mean shift algorithm in Hilbert space

(3) a scalable stochastic variant based on random partitioning, with partial theoretical justification

(4) real-data applications demonstrating the method's scalability and practical usefulness

To ensure consistency in the online technical program, this abstract template must be used for submissions.

Keywords: mean shift clustering, functional data analysis, big data, convergence analysis, randomized algorithm

Personalized Functional Principal Component Analysis with Applications

Ruey S. Tsay[1,2], Ming-Chung Chang[3], Yen-Shiu Chin[3]

[1]*Booth School of Business, University of Chicago, Chicago, IL, USA*

[2]*Institute of Statistics and Data Science, National Tsing-Hua University, Hsinchu, Taiwan*

[3]*Institute of Statistical Science, Academia Sinica, Taipei, Taiwan*

ABSTRACT

Large spatio-temporal functional data are widely available nowadays. They are important in investigating global warming and climate changes. Dimension reduction becomes essential in studying those large functional data and functional principal component analysis is one of the commonly used dimensional reduction methods. On the other hand, similarly to other big data, large spatio-temporal functional data often exhibit certain commonality and some specific local features. The conventional functional principal component analysis becomes inadequate to handle such heterogeneous functional data. In this talk, we generalize personalized PCA to personalized functional PCA. We address both the computational and theoretical issues. For applications, we apply the proposed personalized functional PCA to the PM2.5 measurements in Taiwan and ARGO data.

Keywords: Functional principal component analysis, Functional data, Climate change, Dimension reduction

Nonstationary Gaussian Scale Mixture Models for Spatial Functional Extremes

Yen-Shiu Chin[1], Hsin-Cheng Huang[1], Nan-Jung Hsu[2]

[1]*Institute of Statistical Science, Academia Sinica*

[2]*Institute of Statistics and Data Science, National Tsing Hua University*

ABSTRACT

Extreme events are rare but highly impactful. For example, the prolonged heavy rainfall lasting five hours on June 20, 2021, in southern Taiwan caused agricultural losses exceeding NTD 75 million. Spatial modeling of such extreme environmental events viewed as realizations of functional processes defined over continuous spatial domains is essential for risk assessment in climatology, hydrology, and related fields.

Gaussian scale mixture (GSM) models, constructed as the product of a standard Gaussian process and a positive random scale, provide great flexibility for modeling spatial extremes. An important property of GSM models is that the tail behavior of the random scale determines their extremal dependence structure. Heavy-tailed scales, such as Pareto- or inverse-Gamma-type behavior, lead to asymptotic dependence through extreme shocks in the mixture and simultaneous occurrences of extreme events, whereas lighter-tailed scales, such as Weibull- or Rayleigh-type distributions, yield asymptotic independence. Despite this flexibility, existing GSM models are generally restricted to stationary or isotropic forms, limiting their applicability to complex environmental phenomena.

In this talk, we propose a new class of nonstationary Gaussian scale mixture models using a basis-function approach. Model parameters are estimated via maximum likelihood, and the computation is implemented efficiently through an expectation-maximization algorithm. We demonstrate the advantages of the proposed method through simulation studies and an environmental application to Irish temperature extremes.
Our findings show that allowing nonstationarity leads to more realistic spatial representations of environmental extremes and provides a modern functional data perspective for modeling spatial risk in environmental sciences.

Keywords: extremal dependence structure, Gaussian scale mixture models, maximum likelihood, spatial random-effects models.

Optimal Subdata Selection for Large-Scale Linear Regression Under Model Misspecification

Yundi Kong, **Min Yang**, and Ping-Shou Zhong
University of Illinois at Chicago

ABSTRACT

Large-scale regression problems frequently arise in real world applications, and its substantial computational cost motivates the need for subdata selection methods which balance estimation efficiency and accuracy. Many existing work and approaches are designed to optimize parameter estimation efficiency but can be sensitive to model misspecification or rely on assumptions that are difficult to justify in practice.

Motivated by the need for methods that remain effective under both well-specified and misspecified models, we develop a new subset selection framework and derive its corresponding objective function. We show that minimizing the expected prediction error in both well-specified and misspecified linear models leads to an A-optimality-type criterion that unifies model-based and prediction-oriented perspectives. To illustrate the practical implications of this result, we instantiate the framework using the existing IBOSS+ algorithm, which selects informative subsets according to our derived criterion. Our empirical studies confirm that the proposed method achieves superior performance over existing methods under model misspecification setting.

Keywords: A-optimality, Coreset, IBOSS

Influence-Guided Active Subsampling for High-Dimensional Ridge Regression with Application in GWAS

Lin Wang

Department of Statistics, Purdue University

ABSTRACT

Despite the availability of extensive data sets, it is often impractical to collect labels for all data points in many applications due to various measurement constraints. Subsampling approaches can be employed to select a subset of design points from a large pool, resulting in substantial savings in experimental costs. However, existing subsampling methods are primarily designed for low-dimensional data or rely on the assumption of sparse significant predictors. In this study, we propose a computationally tractable sampling method that enables the selection of a small subset from a large data set without assuming sparsity. Our method acknowledges the possibility that the number of significant predictors can be as large as or even larger than the sample size of the full data set. Specifically, our focus lies on ridge regression, for which we develop sampling probabilities that minimize the mean squared predictive risk on the full data set. The efficacy of our proposed approach is substantiated through theoretical analysis and extensive simulations. The results demonstrate its superiority over existing subsampling methods when dealing with high-dimensional data containing numerous significant predictors. Additionally, we illustrate the advantages of our new approach through its application to genome-wide association studies, highlighting its potential to yield valuable insights in this domain.

Keywords: Experimental design; Active learning

Robust Data Fusion via Subsampling

HaiYing Wang[1], Jing Wang, Kun Chen

Department of Statistics, University of Connecticut

ABSTRACT

Data fusion and transfer learning are rapidly growing fields that enhance model performance for a target population by leveraging other related data sources or tasks. The challenges lie in the various potential heterogeneities between the target and external data, as well as various practical concerns that prevent a naïve data integration. We consider a realistic scenario where the target data is limited in size while the external data is large but contaminated with outliers; such data contamination, along with other computational and operational constraints, necessitates proper selection or subsampling of the external data for transfer learning. To our knowledge, transfer learning and subsampling under data contamination have not been thoroughly investigated. We address this gap by studying various transfer learning methods with subsamples of the external data, accounting for outliers deviating from the underlying true model due to arbitrary mean shifts. Two subsampling strategies are investigated: one aimed at reducing biases and the other at minimizing variances. Approaches to combine these strategies are also introduced to enhance the performance of the estimators. We provide non-asymptotic error bounds for the transfer learning estimators, clarifying the roles of sample sizes, signal strength, sampling rates, magnitude of outliers, and tail behaviors of model error distributions, among other factors. Extensive simulations show the superior performance of the proposed methods. Additionally, we apply our methods to analyze the risk of hard landings in A380 airplanes by utilizing data from other airplane types, demonstrating that robust transfer learning can improve estimation efficiency for relatively rare airplane types with the help of data from other types of airplanes.

Keywords: Mean shifts; Non-asymptotic error bounds; Outliers; Transfer learning

Efficient Subdata Selection for Parameter Estimation

Min Yang[1], **John Stufken[2]**, Ming-Chung Chang[3]

[1]*Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago*

[2]*Department of Statistics, George Mason University*

[3]*Institute of Statistical Science Academia Sinica*

ABSTRACT

When, in terms of the number of data points, the size of a dataset exceeds available computing resources, or when labeling is expensive, an attractive solution consists of selecting only some of the data points (subdata) for further consideration. A central question for selecting subdata of size n from N available data points is which n points to select. While an answer to this question depends on the objective, one approach for a parametric model and a focus on parameter estimation is to select subdata that retains maximal information. Identifying such subdata is a classical NP-hard problem due to its inherent discreteness. Based on optimal approximate design theory, we propose a new methodology for information-based subdata selection, resulting in subdata that approaches the optimal solution. The proposed method is supported by a novel algorithm that is proven to converge and that applies to a general model, accommodates arbitrary choices of N and n, and supports multiple optimality criteria. This is joint work with Min Yang (University of Illinois Chicago) and Ming-Chung Chang (Academia Sinica).

Keywords: Approximate design, Equivalence theorem, Exact design, IBOSS, Optimal design

Advances in Spatial Integer-Valued Time Series Modeling

Cathy W.S. Chen[1], Chun-Shu Chen[2], Hsiao-Hsuan Liao[1]

[1]*Department of Statistics, Feng Chia University, Taichung, Taiwan*

[2]*Graduate Institute of Statistics, National Central University, Taoyuan, Taiwan*

ABSTRACT

This study compares spatial hurdle and spatial zero-inflated generalized Poisson (ZIGP) INGARCH models for analyzing weekly dengue fever counts. Both models extend the INGARCH framework to accommodate spatio-temporal dependence and excess zeros. To enhance epidemiological relevance, we incorporate seasonal components into the log-intensity equations, comparing two approaches: one using periodic harmonic terms based on Fourier series, and the other using meteorological covariates. Model inference is conducted within a Bayesian framework. The spatial hurdle model is further improved using an empirical Bayes approach. Model performance is evaluated using the Deviance Information Criterion (DIC), Bayes factors, and predictive accuracy metrics, including mean squared error (MSE), mean absolute error (MAE), and mean absolute scaled error (MASE). The results highlight the distinct roles of seasonal and environmental covariates in dengue transmission dynamics. Modeling periodic effects with Fourier terms proves effective, especially when explicit meteorological data are unavailable or incomplete. The empirical Bayes approach enhances parsimony and stability over traditional hurdle INGARCH models.

Keywords: Bayesian inference; Fourier series; INGARCH models; Spatio-temporal modelling; Zero-inflated count data

Granger Causality Tests for High-Dimensional VAR Processes

Wei Biao Wu[1], Wentao Hu[1], **Ying-Chao Hung**[2]

[1]*Department of Statistics, University of Chicago*

[2]*Institute of Industrial Engineering, National Taiwan University*

ABSTRACT

Granger causality is a classical and widely used tool for assessing the predictive influence of one group of time series on another within a vector autoregressive (VAR) framework. Traditional Granger causality tests are typically based on Wald-type statistics, but their implementation can be problematic in high-dimensional VAR models due to (i) inflation of the test statistic caused by singular or nearly singular covariance matrices, and (ii) infeasibility or high computational cost associated with tuning parameter selection. In this talk, we introduce an alternative testing procedure for Granger causality that is built on non-pivotal statistics. The proposed method has a solid theoretical foundation and, importantly, does not require any tuning parameter calibration. We further extend the procedure to more general VAR processes with a potentially infinite number of variables, and we establish the corresponding asymptotic theory when the dimensionality of the process increases with the sample size.

Keywords: Vector Autoregression; High-Dimensional Granger Causality; Wald-Type Tests; Non-pivotal Statistics; Gaussian Approximation

Autotune: Fast, Efficient, and Automatic Tuning Parameter Selection for LASSO

Sumanta Basu, Tathagata Sadhukhan

Department of Statistics and Data Science, Cornell University

ABSTRACT

Least absolute shrinkage and selection operator (LASSO) is a popular method for high-dimensional regression, with applications in the analysis of large-scale dependent data such as vector autoregressive (VAR) models in time series. Despite the availability of fast software, interpretability, and asymptotic theory, some practical concerns remain around its tuning parameter selection. Cross-validation (CV), the most common choice in practice, is computationally expensive and comes at the cost of efficiency loss, an issue that is exacerbated for time series cross-validations (TS-CV). Information criteria based tuning is also known to suffer in high-dimensional scenarios. We propose autotune, a procedure that alternately optimizes a penalized log-likelihood over regression coefficients and the error standard deviation, resulting in a LASSO that automatically tunes itself. The premise of autotune is that under exact or approximate sparsity conditions, the error standard deviation may be estimated more easily than the high-dimensional regression parameter. We achieve this algorithmically by leveraging the partial residuals (PR) that are already computed when finding a LASSO solution using coordinate descent. Using extensive simulation experiments on regression and VAR estimation, we show that autotune is faster, and provides superior estimation, variable selection, and prediction performance than existing tuning strategies for LASSO as well as alternatives such as the scaled LASSO. As a by product, autotune provides a new estimator or σ that can be used for high-dimensional inference. Using the partial residuals, we also propose a new visual diagnostic procedure for checking the sparsity assumption. Finally, we demonstrate the utility of autotune on a real-world financial data set. An R package based on C++ is also available on Github.

Keywords: LASSO; Vector Autoregression (VAR); Tuning parameter selection; Cross-validation

Improving Inference and Variable Selection for Two-Phase Studies with High-Dimensional Covariates

Haoyang Wang[1], Qingning Zhou[2], and **Kin Yau Wong**[1]

[1]*The Hong Kong Polytechnic University*

[2]*The University of North Carolina at Charlotte*

ABSTRACT

The two-phase study design is widely used to improve estimation efficiency and reduce cost. In many two-phase studies, the outcome and inexpensive covariates are obtained on all subjects in Phase I, while expensive covariates are measured only on a subset of subjects in Phase II. As a result, regression analysis of two-phase studies faces a missing data problem. In this presentation, we consider two-phase studies with high-dimensional covariates, where one faces a more challenging high-dimensional missing data problem. For this problem, complete-case analyses are generally inefficient, while imputation or likelihood-based methods usually require a model for the missing covariates, which is almost impossible to correctly specify in high-dimensional settings. To overcome these limitations, we propose a two-step estimation method that refines a complete-data estimator by incorporating the incomplete data, and auxiliary information if available, to improve estimation efficiency. This method avoids the need to correctly specify a model for the missing covariates, and the resulting estimator is guaranteed to be at least as efficient as the complete-data estimator. We also establish theoretical guarantees for the proposed method, including estimation consistency, inference validity, and variable selection consistency. We evaluate the performance of the proposed method via simulation studies and provide an application to a major cancer study for illustration.

Keywords: Debiased estimation; High-dimensional regression; Missing data; Robust estimation.

Efficient Estimation for Functional Accelerated Failure Time Model

Changyu Liu, Wen Su, **Kin Yat Liu**, Guosheng Yin, Xingqiu Zhao

Department of Statistics and Data Science, The Chinese University of Hong Kong

ABSTRACT

We propose a functional accelerated failure time model to characterize effects of both functional and scalar covariates on the time to event of interest, and provide regularity conditions to guarantee model identifiability. For efficient estimation of model parameters, we develop a sieve maximum likelihood approach where parametric and nonparametric coefficients are bundled with an unknown baseline hazard function in the likelihood function. Not only do the bundled parameters cause immense numerical difficulties, but they also result in new challenges in theoretical development. By developing a general theoretical framework, we overcome the challenges arising from the bundled parameters and derive the convergence rate of the proposed estimator. Furthermore, we prove that the finite-dimensional estimator is root-n consistent, asymptotically normal and achieves the semiparametric information bound. And we demonstrate the nonparameteric optimality of functional estimator and construct the asymptotic simultaneous confidence band. The proposed inference procedures are evaluated by extensive simulation studies and illustrated with an application to the National Health and Nutrition Examination Survey data.

Keywords: Functional Data Analysis; Survival Analysis

Semiparametric Causal Inference for Right-Censored Outcomes with Many Weak Invalid Instruments

Qiushi Bu[1,2], Wen Su[1], Xingqiu Zhao[3], and Zhonghua Liu[4]

[1]*Department of Biostatistics, City University of Hong Kong, Hong Kong*

[2]*Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China*

[3]*Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong*

[4]*Department of Biostatistics, Columbia University, New York, NY, USA*

ABSTRACT

We propose a semiparametric framework for causal inference with right-censored survival outcomes and many weak invalid instruments, motivated by Mendelian randomization in biobank studies where classical methods may fail. We adopt an accelerated failure time model and construct a moment condition based on augmented inverse probability of censoring weighting, incorporating both uncensored and censored observations. Under a heteroscedasticity-based condition on the treatment model, we establish point identification of the causal effect despite censoring and invalid instruments. We propose GEL-NOW (Generalized Empirical Likelihood with Non-Orthogonal and Weak moments) for valid inference under these conditions. A divergent number of Neyman orthogonal nuisance functions is estimated using deep neural networks. A key challenge is that the conditional censoring distribution is a non-Neyman orthogonal nuisance, contributing to the first-order asymptotics of the estimator for the target causal effect parameter. We derive the asymptotic distribution and explicitly incorporate this additional uncertainty into the asymptotic variance formula. We also introduce a censoring-adjusted over-identification test that accounts for this variance component. Simulation studies and UK Biobank applications demonstrate the method's robustness and practical utility.

Keywords: Censored outcomes; Deep neural networks; Generalized empirical likelihood; Mendelian randomization; Over-identification test; Semiparametric theory; Weak and invalid instruments

Efficient Estimation for Deep Accelerated Failure Time Model with Application to Credit Risk Analysis

Kun Ren[1], Li Liu[2,*], Wen Su[1], Xingqiu Zhao[3,*]

[1]*Department of Biostatistics, City University of Hong Kong*

[2]*School of Mathematics and Statistics, Wuhan University*

[3]*Department of Applied Mathematics, The Hong Kong Polytechnic University*

ABSTRACT

Time-to-event data, frequently encountered in credit risk analysis and engineering reliability studies, present significant analytical challenges due to censoring and data heterogeneity. We develop a deep semiparametric accelerated failure time model designed for right-censored data, which accommodates complex data structures while maintaining interpretability through parametric components. Our approach integrates the flexibility of deep neural networks with traditional smoothing techniques, preserving the benefits of parametric interpretability. Furthermore, we establish both nonasymptotic and asymptotic properties for the proposed method, including prediction error bounds, asymptotic normality, and semiparametric efficiency of the estimator. Simulation studies demonstrate the superior performance of our method compared to conventional smoothing approaches. Finally, we apply our method to analyze the German credit data.

Keywords: Default risk; Deep neural network; Semiparametric efficiency; Survival data.

Nonparametric GARCH: A Deep Learning Approach

Ruizhi Deng[1], **Guohao Shen**[1] and Ngai Hang Chan[2]

[1]*Department of Applied Mathematics, The Hong Kong Polytechnic University*

[2]*Department of Biostatistics, The City University of Hong Kong*

ABSTRACT

This paper introduces a novel approach to estimating nonparametric GARCH models using deep neural networks. We propose an efficient iterative algorithm for training these deep estimators, characterized by ease of implementation and adaptability to various model settings and loss functions. Under mild conditions, we establish learning guarantees for the proposed method by deriving non-asymptotic upper bounds on the prediction error. Specifically, we analyze the dependence structure of the iterative learning algorithm with time-series data and derive the orthogonality property for least squares estimation with dependent data, which may hold independent theoretical interest. We also demonstrate that our deep neural network estimator can adapt to the true lag dimension of the volatility model even when the input dimension is overspecified. This crucial property ensures optimal performance even with suboptimal input choices. Through extensive simulations, we validate the effectiveness of our approach, showcasing its superiority over competing methods, particularly in high-dimensional, nonlinear, and complex volatility scenarios. We further demonstrate the practical utility of our deep nonparametric GARCH estimator by applying it to real-world financial data.

Keywords: Deep Neural Network; GARCH model; Nonparametric Methods; Statistical Learning Theory; Volatility estimation

Learning Summary Statistic for Likelihood-Free Inference

Rong Tang, Rui Li

Hong Kong University of Science and Technology, Department of Mathematics

ABSTRACT

The challenge of performing Bayesian inference in models where likelihood functions are difficult to evaluate but sampling is straightforward has driven the development of likelihood-free methods such as approximate Bayesian computation (ABC). ABC circumvents the need for an explicit likelihood by using a rejection sampling approach to approximate the posterior distributions of parameters, which involves accepting or rejecting proposed parameter values based on their ability to generate simulated data that closely resemble the observed data. A key element in ABC is the use of summary statistics to reduce the dimensionality of the data, thereby avoiding the curse of dimensionality in nonparametric conditional density estimation as the observed data size grows. However, the selection of informative summary statistics that can capture the essential information about the parameter contained in the full data remains challenging. This work proposes a general framework for learning informative summary statistics and the subsequent posterior inference based on the summaries. The proposed method provides a global posterior approximation applicable to any dataset. In addition, a more refined posterior approximations for specific datasets can be obtained by integrating this approach with MCMC-ABC methods.

Keywords: Summary Statistics, Approximate Bayesian Computation, Likelihood-free Inference

A Framework for Comprehensive Model and Variable Selection

Xiyue Liao[1], Mary C Meyer[2,*]

[1]*Department of Mathematics and Statistics, San Diego State University*

[2,*]*Department of Statistics, Colorado State University*

ABSTRACT

We propose a framework for choosing variables and relationships without assuming additivity or parametric forms. The relationships between the response and each of the continuous predictors are modelled with regression splines and assumed to be smooth and one of the following: increasing, decreasing, convex, concave, or a combination of monotonicity and convexity. The eight shapes include a wide range of popular parametric functions such as linear, quadratic, exponential, etc., and the set of choices is appropriate if the component functions "do not wiggle." An ordinal predictor can have its set of possible orderings, such as increasing, decreasing, tree or umbrella orderings, no ordering, or constant. Interactions between continuous predictors will be modelled as multi-dimensional warped-plane spline surfaces, where the same possibilities for shapes are considered. We propose combining stepwise selection methods with information criteria, LASSO ideas, and model selection using a genetic algorithm.

Keywords: variable selection; shape and order constraints; nonparametric

Limiting Spectral Distribution of Stochastic Block Model

Mei-Hui Guo

(Joint work with Giap Van Su, May-Ru Chen, and Hao-Wei Huang)

Department of Applied Mathematics, National Sun Yat-sen University, Kaohsiung, Taiwan

ABSTRACT

The stochastic block model (SBM) is an extension of the Erd˝os–R´enyi graph and has applications in numerous fields, such as data analysis, recovering community structure in graph data and social networks. In this paper, we consider the normal central SBM adjacency matrix with K communities of arbitrary sizes. We derive an explicit formula for the limiting empirical spectral density function when the size of the matrix tends to infinity. We also obtain an upper bound for the operator norm of such random matrices by means of the Stieltjes transform and random matrix theory.

Keywords: Stochastic block model; clustering; semi-circular law; spectral distribution.

On the Limiting Properties of Empirical Spectral Distributions in Community-structured Networks

May-Ru Chen

Department of Applied Mathematics, National Sun Yat-sen University

ABSTRACT

The stochastic block model (SBM) generalizes the Erdős-Rényi graph by partitioning nodes into distinct subsets known as blocks or communities. In this talk, we will briefly review the foundational framework for analyzing the convergence of the empirical spectral distribution (ESD) of large-dimensional random matrices as their dimensions grow. Then we demonstrate that, under certain conditions, the ESD converges in a different sense. Additionally, I will discuss the asymptotic behavior of the ESD of random matrices and examine the limiting ESD of relative SBMs.

Keywords: random matrix, stochastic block model, empirical spectral distribution

Hsiau, Shoou-Ren

Recent Initiatives of PMDA: Highlights and Selected Review Cases

Takumi Aoki[1]

Biostatistics Group, Center for product Evaluation, Pharmaceuticals and Medical Devices Agency (PMDA)

ABSTRACT

The Pharmaceuticals and Medical Devices Agency (PMDA) is Japan's regulatory agency for pharmaceuticals and medical devices. It has three main responsibilities: scientific reviews of marketing authorization application of pharmaceuticals and medical devices, monitoring of their post-marketing safety, and providing relief compensation for sufferers from adverse drug reaction and infections by pharmaceuticals or biological products. The PMDA is the only public agency in the world that combines these three functions.

In this presentation, first, the recent initiatives within the PMDA's Biostatistics Group will be introduced. This presentation will outline the electronic data submission, Data Science Round Table Discussion (DSRT) and the recently issued Early Consideration “Points to consider for externally controlled trials”. The electronic data submission, which means submission of subject-level clinical study data in electronic format with new drug submission, is expected to lead to cross-product research and the associated streamlining of clinical development by PMDA. DSRT is an annual workshop for young biostatisticians from industry, academia, and PMDA to discuss practical topics. This year, the discussion topics were “estimand”, “digital biomarker”, and “innovative trial design in practice”. The “Points to consider for externally controlled trials” summarizes PMDA's current thinking about externally controlled trials using subject-level data from either RWD or clinical trials as external control of a single-arm trial.

Next, advanced designs and related review issues we have encountered during actual new drug review will be presented. The examples include Vutrisiran for transthyretin-mediated familial amyloidotic polyneuropathy and Efgartigimod Alfa (Genetical Recombination) for chronic inflammatory demyelinating polyneuropathy, approved in Japan in 2022 and 2024, respectively. For Vutrisiran, an externally controlled trial using data from previous clinical trials was conducted, and the adequacy of the comparison with external control was discussed. For Efgartigimod Alfa, a randomized discontinuation design was adopted, and its effectiveness in patients with active disease was discussed.

Keywords: External controlled trial; Randomized discontinuation design; New drug review

Statistical Methods and Regulatory Considerations in Phase I Clinical Trials

Tzy-Chy Lin

Division of New Drugs, Center for Drug Evaluation, Taipei, Taiwan

ABSTRACT

This presentation explores the statistical methodologies and regulatory considerations pertinent to Phase I clinical trials. It begins with an introduction to the fundamental objectives of Phase I trials, which primarily aim to assess the safety, dosage, and pharmacokinetics of investigational drugs. The discussion then delves into various dose-escalation designs, including traditional, model-based and model-assisted approaches, highlighting their advantages and limitations in determining the maximum tolerated dose (MTD).

Subsequently, the presentation examines the role of simulation studies in optimizing trial designs and predicting outcomes. A hypothetical trial is presented to illustrate the application of these statistical methods in a real-world context. The session concludes with remarks on the importance of adhering to regulatory guidelines, such as those from the International Council for Harmonization (ICH), to ensure the ethical conduct and scientific validity of clinical trials.

Key words: Phase I clinical trials, Dose-escalation designs; Maximum tolerated dose (MTD)

Adaptive Designs for Clinical Trials: Principles and Recommendations from the Draft ICH E20 Guideline

Frank Bretz

Novartis Pharma AG

ABSTRACT

The International Council for Harmonisation (ICH) has released its draft E20 guideline on "Adaptive Designs for Clinical Trials" for public consultation. It acknowledges the high potential for adaptive designs to accelerate the process of drug development and to allocate resources more efficiently without lowering scientific and regulatory standards. Some of the approaches may affect the nature and timing of interactions between industry and regulators at confirmatory trial planning and assessment. The final guideline will indicate key adaptive design principles and approaches for which discussion of adaptive design features, and the rationale for their use, are particularly critical at the planning stage.

This presentation reviews the draft ICH E20 guideline and outlines a transparent and harmonised set of principles for the design, conduct, analysis, and interpretation of adaptive clinical trials, with the aim of supporting regulatory review of these studies in global drug development programmes. These principles are also intended to provide flexibility for evaluating and discussing innovative approaches to clinical trial design throughout the development process.

Keywords: Adaptive design; ICH

Evolving Regulatory Statistical Considerations in Drug Development and Evaluation

Helen Li

ABSTRACT

Statistical methodologies in drug development and regulatory evaluation have evolved significantly, driven by advances in trial design and regulatory expectations. This presentation reviews the history of statistical methods in regulatory submissions, highlights recently adopted approaches like adaptive designs, imputations, estimands, and Bayesian statistics, and explores future trends shaped by the integration of artificial intelligence (AI) in regulatory agencies. By examining past and present practices, we project how AI-driven analytics and real-world evidence will transform statistical practices in drug development.

Deep Kernel Aalen-Johansen Estimator: An Interpretable and Flexible Neural Net Framework for Competing Risks

Xiaobin Shen[*], **George H. Chen**[*,†]

Heinz College of Information Systems and Public Policy, Carnegie Mellon University

[*] *equal contribution*

[†] *presenting author*

ABSTRACT

We propose an interpretable deep competing risks model called the Deep Kernel Aalen-Johansen (DKAJ) estimator, which generalizes the classical Aalen-Johansen nonparametric estimate of cumulative incidence functions (CIFs). Each data point (e.g., patient) is represented as a weighted combination of clusters. If a data point has nonzero weight only for one cluster, then its predicted CIFs correspond to those of the classical Aalen-Johansen estimator restricted to data points from that cluster. These weights come from an automatically learned kernel function that measures how similar any two data points are. On four standard competing risks datasets, we show that DKAJ is competitive with state-of-the-art baselines (that are not interpretable) while being able to provide visualizations to assist model interpretation.

Keywords: survival analysis, competing risks, neural networks, interpretability

Quantifying Uncertainty in Answers from any Language Model and Enhancing their Trustworthiness

Jonas Mueller

Cleanlab

ABSTRACT

We introduce methods for detecting bad and speculative answers from a pretrained Large Language Model by estimating a numeric confidence score for any output it generated. Our uncertainty quantification techniques work for any LLM accessible only via a black-box API, whose training data remains unknown. Experiments on both closed and open-form Question-Answer benchmarks reveal that our approach more accurately identifies incorrect LLM responses than alternative uncertainty estimation procedures (across many frontier models). By sampling multiple responses from the LLM and considering the one with the highest confidence score, we can additionally obtain more accurate responses from the same LLM, without any extra training steps.

Keywords: Large Language Model; Uncertainty Quantification; Trustworthy AI

Accessing the Impact of Data Alteration: When Can R-Squared from Synthetic (or Corrupted) Data Be Trusted?

Xiao-Li Meng, James Bailie, Mohammad Kakooei, Adel Daoud

Department of Statistics, Harvard University

ABSTRACT

Motivated by a seemingly counterintuitive finding in a poverty study that linked privacy-protected household data with satellite imagery via deep learning, we set out—and invite readers to join—to explore the mathematical relationships between an algorithm's performance on two data sets: the original, intended data and an altered version of it. We propose a comparative framework that is agnostic to the alteration mechanism, whether the modification reflects beneficent intent (e.g., synthetic data for privacy protection) or maleficent intervention (e.g., corruption or distortion). This agnostic stance directs attention to the inherent complexity of a "three-body" problem: the relationships between the unaltered and altered data sets, and between how an algorithm interacts with each. To analyze this complexity in a general and coherent way, we introduce the neologism *data alterity*—a notion extending beyond syntheticity—and decompose it qualitatively into *target alterity* and *residual alterity* for a given predictive algorithm and evaluative metric.

The commonly used metric R-squared provides a low-hanging fruit for demonstrating the potential of this framework, through an algebraic relationship linking the unaltered R-squared, the altered R-squared, and an alterity-impact score. This relationship reveals how alterity can help or harm performance by affecting target and residual components differently, regardless of whether the alteration is beneficent or maleficent. We further show that the altered R-squared is typically conservative when alteration behaves like independent noise added to the target variable, as in differential privacy, and we provide a computable adjustment to correct for this case. We also identify a necessary and sufficient condition under which residual alterity exceeds one, implying conservativeness of the altered R-squared whenever the target variable remains unaltered. These initial insights are illustrated in a proxy study predicting ground-level features from Earth observations whose locations have been synthetically perturbed for privacy protection.

Keywords: Alterity-Impact Score; Data Alterity; Data Privacy; Deep Learning; Digital Twins; Residual Alterity; Target Alterity.

On the Asymptotic Properties of Product-PCA under the High-Dimensional Setting

Hung Hung

Institute of Health Data Analytics and Statistics, National Taiwan University, Taiwan

ABSTRACT

Principal component analysis (PCA) is a widely used dimension reduction method, but is known to be non-robust in the presence of outliers. Recently, product-PCA (PPCA) has been shown to possess the efficiency-loss free ordering-robustness property: (i) in the absence of outliers, PPCA and PCA share the same asymptotic distributions; (ii) in the presence of outliers, PPCA is more ordering-robust than PCA in estimating the leading eigen-subspace. This property makes PPCA different from the branch of the conventional robust PCA methods (which usually suffer the problem of efficiency loss in an exchange of robustness gain), and may deserve further investigations. In this article, we study the high-dimensional asymptotic properties of the PPCA eigenvalues via the techniques of random matrix theory. In particular, we derive the critical value for being distant spiked eigenvalues, the limiting values of the sample spiked eigenvalues, and the limiting spectral distribution of PPCA. These results enable us to more clearly observe the superiorities of PPCA in comparison with PCA. Similar to the case of PCA, the explicit forms of the asymptotic properties of PPCA become available under the special case of the simple spiked model. Numerical studies are conducted to verify our results.

Keywords: efficiency-loss; PCA; random matrix theory; robustness

Automatic Sparse Estimation of High-Dimensional Covariance Matrices

Tetsuya Umino[1], **Kazuyoshi Yata[2]**, Makoto Aoshima[2]

[1] *Degree Programs in Pure and Applied Sciences, Graduate School of Science and Technology, University of Tsukuba*

[2] *Institute of Mathematics, University of Tsukuba*

ABSTRACT

Sparse principal component analysis (SPCA) has been widely investigated as a framework for estimating sparse principal component directions in high-dimensional data. A recent study by Yata and Aoshima (Statistica Sinica, 35) proposed the automatic SPCA (A-SPCA), which determines its threshold adaptively and has been shown to achieve consistency under mild conditions. Motivated by this development, we extend the A-SPCA methodology to the estimation of high-dimensional covariance matrices as well as mean vectors, and we introduce novel automatic sparse estimators within this framework. The proposed estimators inherit the adaptability of A-SPCA and are theoretically proven to be consistent under mild assumptions in high-dimensional settings, including challenging high-dimension, low-sample-size contexts. Their performance is carefully examined through numerical simulations, which demonstrate both accuracy and stability across a range of scenarios. In addition, we extend the approach to high-dimensional regression by employing the proposed estimators. Applications to gene expression data confirm their effectiveness in practice.

Keywords: Cross-covariance matrix; Extended cross-data-matrix methodology; Large p small n; High-dimensional regression

Collaborative and Federated Black-Box Optimization: A Bayesian Optimization Perspective

Raed Al Kontar*

Industrial & Operations Engineering, University of Michigan

ABSTRACT

We focus on collaborative and federated black-box optimization (BBOpt), where agents optimize their heterogeneous black-box functions through collaborative sequential experimentation. From a Bayesian optimization perspective, we address the fundamental challenges of distributed experimentation, heterogeneity, and privacy within BBOpt, and propose three unifying frameworks to tackle these issues: (i) a global framework where experiments are centrally coordinated, (ii) a local framework that allows agents to make decisions conditioned on shared information, and (iii) a predictive framework that enhances local surrogates through collaboration to improve decision-making. We categorize existing methods within these frameworks and highlight key open questions to unlock the full potential of federated BBOpt. Our overarching goal is to shift federated learning from its predominantly descriptive/predictive paradigm to a prescriptive one, particularly in the context of BBOpt — an inherently sequential decision-making problem.

Keywords: Federated Leaning, Personalization, Bayesian Optimization, Experimental Design, Distributed Learning

Effective Permutation Tests for Differences Across Multiple High-Dimensional Correlation Matrices

José A. Sánchez Gómez[1], **Elio Zhang**[2], Yufeng Liu*

[1] *Department of Statistics, University of California at Riverside*

[2] *Department of Mathematics, Seattle University*

* *Department of Statistics, University of Michigan*

ABSTRACT

Testing the equality of two or multiple correlation or covariance matrices is an important problem in biology, finance and many other areas. High dimensionality, where the number of features is much larger than the sample size, causes conventional procedures to perform poorly, as they are often based on limiting distributions of test statistics in the classical large sample size setting. Moreover, their performance is often contingent on whether the matrix of differences is sparse or dense, while such information is rarely available. In this article, we develop a new family of permutation testing procedures to tackle these challenges. The introduced tests are demonstrated to outperform many other competing procedures in terms of size control and power under various settings. In particular, using our variance-stabilizing transformation, the proposed methods provide the best performance for testing correlation or covariance matrix differences in both sparse and dense settings. We establish non-asymptotic guarantees on the power of our test, which ensure its reliability for sparse and dense differential correlation matrices. Through the analysis of gene-expression and brain imaging data, we showcase the high power and accurate size control of our test in high-dimensional statistical applications.

Keywords: Brain Activation; Concentration Inequalities; Covariance Testing; Gene Expression; High-Dimensional Statistics

On Model Selection for Causal Inference

CY (Chor-yiu) SIN

College of Technology Management, National Tsing Hua University

ABSTRACT

Causal inference in financial econometrics becomes one of the hot topics in the literature. See, for instance, Liu, Masulis and Stanfield (2021, *JFE*), Carlin, Umar and Yi (2023, *JFE*), and Dasgupta, Huynh and Xia (2023, *RFS*, 2023). In this paper, we modify the covariate selection criterion (CSC) developed by Lu (2015, *JBES*), and minimize the mean squared errors (MSE) of the *focused* parameter in causal methods such as regression discontinuity design (RDD) and difference-in-differences (DiD). Synthetic controls (SC) and inverse probability weighting (IPW) estimator will also be discussed. Monte-Carlo simulations and empirical examples are performed. Our modified CSC are compared with the usual model selection criteria such as AIC, Lasso and OGA+HDIC. We also discuss the further extensions to: (i) partially linear models; and (ii) high-dimensional models.

Keywords: Covariate selection criterion (CSC); difference-in-differences (DiD); focused parameter; mean squared errors (MSE); regression discontinuity design (RDD)

Multi-View Dynamic Network Modeling

Mike So

The Hong Kong University of Science and Technology

ABSTRACT

A flexible multi-view dynamic network model is developed using a regression-like structure, incorporating exogenous and endogenous variables from the lagged networks to model edge changes. The model does not rely on latent space, simplifying network estimation and prediction. Furthermore, it integrates a multi-view feature to represent various relationship types at each time point. The proposed model offers an intuitive interpretation of the estimation. Bayesian model averaging method is also applied to predict networks.

Keywords: Bayesian analysis, network modeling, time series analysis

Probabilistic Loss Reserving Prediction via Denoising Diffusion Model

Shiying Gao, Yuning Zhang, Ruikun Li, **Boris Choy**, Junbin Gao

Discipline of Business Analytics, The University of Sydney, Australia

ABSTRACT

This paper introduces an innovative approach to predicting loss reserves in the insurance industry through a revised diffusion model. This model leverages run-off triangles of claim data as graphical representations, highlighting the interconnections among data points within the triangle. Unlike the traditional cross-classified over-dispersed Poisson (ccODP) model, our proposed diffusion model not only enhances accuracy and efficiency but also provides probabilistic forecasts. Through comprehensive simulation and empirical studies, we demonstrate the superior forecasting capabilities of our diffusion model compared to existing methods. These findings indicate that using network-based interactions within run-off triangles can significantly improve loss reserve forecasting.

Keywords: Machine Learning; Diffusion Model; Loss Reserving; Run-off Triangle

Adaptive Debiased Lasso in High-dimensional Generalized Linear Models with Streaming Data

Ruijian Han[1,*], Lan Luo[2,*], **Yuanhang Luo[1,*]**, Yuanyuan Lin[3], Jian Huang[1,4]

[1]*Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University*

[2]*Department of Biostatistics and Epidemiology, Rutgers School of Public Health*

[3]*Department of Statistics, The Chinese University of Hong Kong*

[4]*Department of Applied Mathematics, The Hong Kong Polytechnic University*

ABSTRACT

Online statistical inference facilitates real-time analysis of sequentially collected data, making it different from traditional methods that rely on static datasets. This paper introduces a novel approach to online inference in high-dimensional generalized linear models, where we update regression coefficient estimates and their standard errors upon each new data arrival. In contrast to existing methods that either require full dataset access or large-dimensional summary statistics storage, our method operates in a single-pass mode, significantly reducing both time and space complexity. The core of our methodological innovation lies in an adaptive stochastic gradient descent algorithm tailored for dynamic objective functions, coupled with a novel online debiasing procedure. This allows us to maintain low-dimensional summary statistics while effectively controlling the optimization error introduced by the dynamically changing loss functions. We establish the asymptotic normality of our proposed Adaptive Debiased Lasso (ADL) estimator. We conduct extensive simulation experiments to show the statistical validity and computational efficiency of our ADL estimator across various settings. Its computational efficiency is further demonstrated via a real data application to the spam email classification.

Keywords: Confidence interval, lasso, one-pass algorithm, stochastic gradient descent

DeepSuM: A Deep Sufficient and Efficient Modality Learning Framework

Zhe Gao[1], Jian Huang[2], **Ting Li[3*]**, Xueqin Wang[1]

[1]*School of Management, University of Science and Technology of China, Hefei, Anhui, People's Republic of China*

[2]*Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong, People's Republic of China*

[3]*Department of Statistics & Data Science, Southern University of Science & Technology, Shenzhen, People's Republic of China*

[*]*Address for correspondence. Ting Li, Southern University of Science & Technology, Shenzhen, 518055, China. liting@sustech.edu.cn*

ABSTRACT

Multimodal learning has become a pivotal approach in developing robust learning models with applications spanning multimedia, robotics, large language models, and healthcare. The efficiency of multimodal systems is a critical concern, given the varying costs and resource demands of different modalities. This underscores the necessity for effective modality selection to balance performance gains against resource expenditures. In this study, we propose a novel framework for modality selection that independently learns the representation of each modality. This approach allows for the assessment of each modality's significance within its unique representation space, enabling the development of tailored encoders and facilitating the joint analysis of modalities with distinct characteristics. Our framework aims to enhance the sufficiency and effectiveness of multimodal learning by optimizing modality integration and selection.

Keywords: Multimodal learning, Representation, Modality integration, Modality selection

Lead Science and Clinical Research – Career Panel Discussion

ABSTRACT

The evolving landscape of pharmaceutical sciences and clinical research offers opportunities for professionals aiming to impact the future of healthcare. This panel discussion brings together thought leaders and senior executives from pharmaceutical companies to share insights into career paths, industry trends, and key competencies in biostatistics, clinical data sciences, and biometrics.

Organized by Haoda Fu, Head of Exploratory Biostatistics at Amgen, the session provides attendees with an opportunity to engage with industry veterans who hold leadership roles. Panelists include Jingyuan Yang, from AbbVie; Dacheng Liu from Boehringer Ingelheim; Thomas Liu from Amgen; Xun Chen from AbbVie; and Zhishen Ye at Gilead Sciences.

Through discussion and interactive dialogue, panelists will offer guidance on career advancement, share experiences from their professional journeys, discuss current and future industry challenges, and outline strategies to build expertise in biostatistics and clinical research. This session is aimed at professionals at various career stages seeking advice, mentorship, and networking opportunities in the field of pharmaceutical science and clinical research.

Organizer: Haoda Fu, Head of Exploratory Biostatistics, Amgen

Panel:

- **Dacheng Liu**, Highly Distinguished Therapeutic Area and Methodology Statistician at Boehringer Ingelheim
- **Jingyuan Yang**, Executive Director, Biostatistics at AbbVie
- **Thomas Liu**, Executive Director, Global Biostatistics Therapeutics Area Head, Center for Design & Analysis at Amgen
- **Xun Chen**, VP, Data and Statistical Sciences, AbbVie
- **Lei Wang**, The Lotus Group, LLC, USA

An Integrated GMM Shrinkage Approach with Consistent Moment Selection from Multiple External Sources

Jun Shao

University of Wisconsin-Madison

ABSTRACT

Interest has grown in analyzing primary internal data by utilizing some independent external aggregated statistics for efficiency gain. However, when population heterogeneity exists, inappropriate incorporation may lead to a biased estimator. With multiple external sources under generalized estimation equations and possibly heterogeneous populations, we propose an integrated generalized moment method that can perform a data-driven selection of valid moment equations from external sources and make efficient parameter estimation simultaneously. Moment equation selection consistency and asymptotic normality are established for the proposed estimator. Further, when the sample sizes of all external sources are large compared to the internal sample size, asymptotically the proposed estimator is more efficient than the estimator based on the internal data only and is oracle-efficient in the sense that it is as efficient as the oracle estimator based on all valid moment equations. Simulation studies confirm the theoretical results and the efficiency of the proposed method empirically. An example is also included for illustration.

Keywords: Adaptive lasso; Data integration; Generalized method of moments; Heterogeneous population

Training-Free Multi-Agent Language Models

Xiaowu Dai

Departments of Statistics and Data Science and of Biostatistics, UCLA

ABSTRACT

Large Language Models (LLMs) have demonstrated strong generative capabilities but remain prone to inconsistencies and hallucinations. We introduce Peer Elicitation Games (PEG), a training-free, game-theoretic framework for aligning LLMs through a peer elicitation mechanism involving a generator and multiple discriminators instantiated from distinct base models. Discriminators interact in a peer evaluation setting, where rewards are computed using a determinant-based mutual information score that provably incentivizes truthful reporting without requiring ground-truth labels. We establish theoretical guarantees showing that each agent, via online learning, achieves sublinear regret in the sense their cumulative performance approaches that of the best fixed truthful strategy in hindsight. Moreover, we prove last-iterate convergence to a truthful Nash equilibrium, ensuring that the actual policies used by agents converge to stable and truthful behavior over time. Empirical evaluations across multiple benchmarks demonstrate significant improvements in factual accuracy. These results position PEG as a practical approach for eliciting truthful behavior from LLMs without supervision or fine-tuning.

Keywords: LLMs; Nash equilibrium; No-regret learning; Incentives

Optimal-PhiBE for Continuous-Time Reinforcement Learning with Discrete-Time Data

Yuhua Zhu[1]

Department of Statistics and Data Science, University of California-Los Angeles

ABSTRACT

This talk addresses continuous-time reinforcement learning (RL) in settings where the system dynamics are governed by a stochastic differential equation but remains unknown, with only discrete-time observations available. We introduce Optimal-PhiBE, an equation that integrates discrete-time information into a PDE, combining the strengths of both RL and PDE formulations. In linear-quadratic control, Optimal-PhiBE can even achieve accurate continuous-time optimal policy with only discrete-time information. In general dynamics, Optimal-PhiBE is less sensitive to reward oscillations, leading to smaller discretization errors. Furthermore, we extend Optimal-PhiBE to higher orders, providing increasingly accurate approximations. At the end of the talk, I will discuss how this technique can be leveraged to generate time-dependent samples and tackle goal-oriented inverse problems.

Keywords: Reinforcement learning; Optimal control; Time-series data

Statistically and Computationally Optimal Estimation and Inference in Common Subspaces

Joshua Agterberg

Department of Statistics, University of Illinois Urbana-Champaign

ABSTRACT

In this talk we investigate the statistical and computational limits for the common subspace model, a model wherein one observes a collection of symmetric low-rank matrices perturbed by noise, where each low-rank matrix shares the same common subspace. First, we propose an estimator based on projected gradient descent initialized via a spectral sum of squared matrices and show that it achieves the optimal sin Theta error under a strong signal-to-noise ratio (SNR) condition, and we further give evidence that this SNR condition is necessary for a polynomial-time estimator to exist. Next, we turn to estimation and inference for the sin Theta distance itself, and we show that our estimator achieves an asymptotically Gaussian distribution with a bias term that vanishes under a strong signal requirement. Based on this limiting result we propose confidence intervals and show that they are minimax optimal, though the resulting confidence intervals require knowledge of the SNR. We then turn to designing adaptive confidence intervals for the sin Theta error, and we show that adaptivity is information-theoretically impossible unless the SNR is sufficiently strong. Consequently, our results unveil a novel phenomenon: despite the SNR being "above" the computational limit for estimation, adaptive statistical inference may still be information-theoretically impossible.

Keywords: networks, spectral methods, multilayer networks

UBSea: A Unified Community Detection Framework

Xiancheng Lin[1], **Hao Chen[1]**

[1]*University of California, Davis*

ABSTRACT

Detecting communities in networks is an important task across many disciplines, including statistics, social science, and engineering. Existing approaches often target one community structure at a time -- assortative in most cases, and to a lesser extent, disassortative or core-periphery -- without offering a simple, unified framework across all cases. We propose UBSea (Unified Bigroups Standardized Edge-count Analysis), a strategic extension of modularity that unifies the detection of all three mixing patterns. UBSea automatically identifies the appropriate structure and applies to both directed and undirected networks. We establish weak and strong consistency results under the stochastic block model (SBM), expanding the class of graphs where consistent recovery can be achieved by efficient algorithms, including asymmetric and core-periphery settings not covered by previous guarantees. Extensive simulations under the SBM and degree-corrected SBM confirm UBSea's high accuracy across diverse scenarios, and applications to benchmark datasets demonstrate that while other methods can perform well in individual regimes, UBSea provides a single framework that performs reliably across assortative, disassortative, and core–periphery networks.

Keywords: Assortative mixing; Disassortative mixing; Core-periphery structure; Consistency and asymptotics

Finding Anomalous Cliques in Inhomogenous Networks using Egonets

Subhankar Bhadra[1], **Srijan Sengupta[2]**

[1]*Pennsylvania State University*

[2]*North Carolina State University*

ABSTRACT

We consider the problem of finding an anomalous clique, i.e., a fully connected subgraph, hidden in a large network. There are two parts to this problem: (1) detection, i.e., determining whether an anomalous clique is present, and (2) identification or localization, i.e., given that an anomalous clique is detected in part 1, determining which vertices of the network constitute the clique. This problem has a number of practical applications, such as financial trading networks, brain networks, and online social networks. A rich literature already exists on the detection problem when restricted to homogeneous Erd˝os–R´enyi random graphs. However, currently, no method exists that can solve the detection and identification/localization problems for inhomogeneous networks in finite time. We propose an inferential tool based on egonets to address this gap. The proposed method is computationally efficient and naturally amenable to parallel computing and easily extends to a wide variety of inhomogeneous network models. We establish the theoretical properties of the proposed method and demonstrate its empirical performance through simulation studies.

Keywords: Random graphs; Clique; Motif detection; Hypothesis testing

Tackling Explosive Likelihood and Non-Identifiability in Beta Mixtures

Tom Tang[1], Mingxing He[2], Daniel J. McDonald[1], **Jiahua Chen**[1]

[1]*Department of Statistics, the University of British Columbia, Canada.*

[2]*Key Laboratory of Applied Statistics, Kunming University of Science and Technology, China.*

ABSTRACT

This work explores statistical inference under beta mixture models. A key challenge with finite beta mixtures is that the likelihood can explode, and the model itself can be non-identifiable. To address the problem of explosive likelihood, we propose a penalized maximum likelihood estimator that adds a penalty term to the log-likelihood. We also introduce a moment-based alternative, which is more computationally efficient. For order selection in the presence of non-identifiability, we investigate order reduction using composite transportation divergence. We demonstrate the proposed methods through simulation studies and apply them to DNA methylation data. All implementations are available in the https://github.com/jhchen-stat-ubc-ca/MixtureInf2.0, an R package MixtureInf2.0 on GitHub.

Keywords: EM algorithm; Machine Learning; Model-based clustering; Optimal transport

Robust Inverse Normal Transformation-Based Tests under Linear Mixed Effects Models

Elif Acar[1], Sonja Friesen[1]

[1]*Department of Mathematics and Statistics, University of Guelph*

ABSTRACT

Many biomedical applications generate vast amounts of data that cannot be managed under human oversight alone, necessitating automated analysis pipelines. These pipelines require robust statistical methods to handle frequent assumption violations in biological settings. Our research is motivated by the International Mouse Phenotyping Consortium (IMPC), which examines mammalian gene function using genetically edited knockout mice. The IMPC employs a linear mixed effects model (LMM) that considers both fixed effects, such as gene, sex, and weight, and random effects that may arise from shared experimental batch or litter. Using simulations based on the IMPC experiment setup, we assess the robustness of the LMM under violations of the normality assumption for error and random effects. We further investigate whether rank-based inverse normal transformations could enhance the detection power of gene effects in these scenarios.

Keywords: Gene effect; Inverse normal transformation; Mixed effects model; Robust inference

Using Statistical Modelling to Inform Public Health Decision Making in Hepatitis B

William W.L. Wong[1]

School of Pharmacy, University of Waterloo

ABSTRACT

Managing chronic hepatitis B (CHB) is difficult because the majority of those infected have clinical silent disease. The asymptomatic nature of CHB means that the disease often remains undiagnosed, leaving its prevalence highly uncertain. This generates significant uncertainty for policy makers in the planning of hepatitis eradication programs to meet the goals set by the WHO. The objective of this work is to establish a statistical framework for the estimation of CHB prevalence and the undiagnosed proportion. A state-transition model describing infection, disease progression and treatment response was mathematically formulated and developed. We then back-calculated the historical prevalence of CHB and the undiagnosed proportion through a calibration process based on a Bayesian Markov chain Monte Carlo (MCMC) algorithm. The algorithm constructs posterior distributions of the historical prevalence of CHB by comparing the model-generated predictions of the annual numbers of health events related to CHB infection and its sequelae against observed calibration targets. These predicted estimates offer crucial insights for policymakers, aiding in the formulation of effective CHB screening and treatment policies.

Keywords: chronic hepatitis B; Markov chain Monte Carlo; prevalence; public health policy

Variable Selection in Mixture Regression Models

Zeny Feng[1], Grace Stelter[2], Lorna Deeth[2], Alysha Cooper[2]

Department of Mathematics and Statistics, University of Guelph

ABSTRACT

The selection of relevant variables through regularization has been widely used in many studies. Finite mixture regressions have also been extensively employed to explore and model the heterogeneous covariate effects on the response. Despite their complementary nature, the issue of variable selection in mixture regression models has received little attention in the literature so far, resulting in limited methodologies available for fitting regularized mixture regression models. In this talk, we introduce a novel algorithm for optimizing regularized mixture regression models with different choices of penalties. Our proposed method is desirable for retaining only relevant covariates in each of the subpopulations and allowing these relevant covariates to have heterogenous effects in different subpopulations. We applied our proposed method to Chiroptera data for relevant and heterogeneous environmental and biological effects on the evolutional development of bat's forearm.

Keywords: Mixture regression models; variable selection; penalized likelihood; optimization

Multi-Dimensional Distributional Reinforcement Learning: A Hilbert Space Embedding Approach

Mehrdad Mohammadi[1], **Qi Zheng[2]**, Ruoqing Zhu[3]

[1]*Department of Statistics, University of Illinois Urbana-Champaign*

[2]*Department of Bioinformatics and Biostatistics, University of Louisville*

[3]*Department of Statistics, University of Illinois Urbana-Champaign*

ABSTRACT

We propose an (offline) distributional reinforcement learning framework (RK-DRL) that leverages Hilbert space embeddings to estimate the multi-dimensional value distribution under a proposed target policy. In our setting the state-action are also multi-dimensional and continuous. By mapping probability measures into a reproducing kernel Hilbert space via kernel mean embeddings, our method replaces Wasserstein metrics with a novel integral probability metric. This enables efficient estimation in multi-dimensional state–action spaces and reward settings, where direct computation of Wasserstein distances is computationally challenging. Theoretically, we establish contraction properties of the distributional Bellman operator under our proposed metric involving the Mat´ern family of kernels and provide uniform con-vergence guarantees. Empirical results demonstrate improved convergence rates and robust off-policy evaluation under mild assumptions, namely, Lipschitz continuity and boundedness for the kernels, highlighting the potential of our embedding-based approaches in complex, real-world decision-making scenarios and risk evaluations.

Keywords: Wasserstein Distance, Reproducing Kernel Hilbert Space, Non-parametric, Matern Kernel

Tuning Parameter Calibration for Prediction in Personalized Medicine

Shih-Ting Huang

Department of Bioinformatics and Biostatistics, University of Louisville

ABSTRACT

Personalized medicine has become an important part of medicine, for instance predicting individual drug responses based on genomic information. However, many current statistical methods are not tailored to this task, because they overlook the individual heterogeneity of patients. In this paper, we look at personalized medicine from a linear regression standpoint. We introduce an alternative version of the ridge estimator and target individuals by establishing a tuning parameter calibration scheme that minimizes prediction errors of individual patients. In stark contrast, classical schemes such as cross-validation minimize prediction errors only on average. We show that our pipeline is optimal in terms of oracle inequalities, fast, and highly effective both in simulations and on real data.

Keywords: Personalized Prediction; Ridge Regression; Precision Medicine; Parameter Calibration

Principal Stratification with U-Statistics under Principal Ignorability

Fan Li[1], Xinyuan Chen[2]

[1]*Department of Biostatistics, Yale School of Public Health, CT, USA*

[2]*Department of Mathematics and Statistics, Mississippi State University, MS, USA*

ABSTRACT

Principal stratification is a popular framework for causal inference in the presence of an intermediate outcome. While the principal average treatment effects have traditionally been the default target of inference, it may not be sufficient when the interest lies in the relative favorability of one potential outcome over the other within the principal stratum. We introduce the principal generalized causal effect estimands, which extend the principal average causal effects to accommodate arbitrary nonlinear contrast functions. Under principal ignorability, we expand the existing theoretical results to a much wider class of causal estimands in the presence of a binary intermediate variable. We develop identification formulas and derive the efficient influence functions of the generalized causal estimands for principal stratification analyses. These efficient influence functions motivate a set of multiply robust estimators and lay the ground for obtaining efficient debiased machine learning estimators via cross-fitting based on U-statistics. The proposed methods are illustrated through simulations and the analysis of a data example.

Keywords: Causal inference, efficient influence function, principal stratification, multiply robust estimation, win ratio, probabilistic index

Using a Two-Parameter Sensitivity Analysis Framework to Efficiently Combine Randomized and Non-randomized Studies

Ruoqi Yu[1], Bikram Karmakar[2], Jessica Vandeleest[3], Eleanor Bimla Schwarz[4]

[1]*Department of Statistics, University of Illinois Urbana-Champaign*

[2]*Department of Statistics, University of Florida*

[3]*California National Primate Research Center, University of California Davis*

[4]*Department of Medicine, University of California San Francisco*

ABSTRACT

Causal inference is vital for informed decision-making across fields such as biomedical research and social sciences. Randomized controlled trials (RCTs) are considered the gold standard for the internal validity of inferences, whereas observational studies (OSs) often provide the opportunity for greater external validity. However, both data sources have inherent limitations preventing their use for broadly valid statistical inferences: RCTs may lack generalizability due to their selective eligibility criterion, and OSs are vulnerable to unobserved confounding. This paper proposes an innovative approach to integrate RCT and OS that borrows the other study's strengths to remedy each study's limitations. The method uses a novel triplet matching algorithm to align RCT and OS samples and a new two-parameter sensitivity analysis framework to quantify internal and external validity biases. This combined approach yields causal estimates that are more robust to hidden biases than OSs alone and provides reliable inferences about the treatment effect in the general population. We apply this method to investigate effects of lactation on maternal health using a small RCT and a long-term observational health records dataset from the California National Primate Research Center. This application demonstrates the practical utility of our approach in generating scientifically sound and actionable causal estimates.

Keywords: Causal inference; Generalizability bias; Matching, Sensitivity analysis, Unmeasured confounding

Robust Sensitivity Analysis via Augmented Percentile Bootstrap under Simultaneous Violations of Unconfoundedness and Overlap

Cui Han[1], **Xinran Li[2]**

[1]*Department of Statistics, University of Illinois at Urbana-Champaign*

[2]*Department of Statistics, University of Chicago*

ABSTRACT

The identification of causal effects in observational studies typically relies on two standard assumptions: unconfoundedness and overlap. However, both assumptions are often questionable in practice: unconfoundedness is inherently untestable, and overlap may fail in the presence of extreme unmeasured confounding. While various approaches have been developed to address unmeasured confounding and extreme propensity scores separately, few methods accommodate simultaneous violations of both assumptions. In this paper, we propose a sensitivity analysis framework that relaxes both unconfoundedness and overlap, building upon the marginal sensitivity model. Specifically, we allow the bound on unmeasured confounding to hold for only a subset of the population, thereby accommodating heterogeneity in confounding and allowing treatment probabilities to be zero or one. Moreover, unlike prior work, our approach does not require bounded outcomes and focuses on overlap-weighted average treatment effects, which are both practically meaningful and robust to non-overlap. We develop computationally efficient methods to obtain worst-case bounds via linear programming, and introduce a novel augmented percentile bootstrap procedure for statistical inference. This bootstrap method handles parameters defined through over-identified estimating equations involving unobserved variables and may be of independent interest. Our work provides a unified and flexible framework for sensitivity analysis under violations of both unconfoundedness and overlap.

Keywords: unconfoundedness; overlap; extreme confounding; overlap weights; parameter augmentation

Power and Sample Size Calculations for Causal Inference with Observational Data

Fan Li[1], Bo Liu[1]

Department of Statistical Science, Duke University

ABSTRACT

This paper investigates the theoretical foundation and develops analytical formulas for sample size and power calculations for causal inference with observational data. By analysing the variance of the inverse probability weighting estimator of the average treatment effect, we decompose the power calculations into three components: propensity score distribution, potential outcome distribution, and their correlation. We show that to determine the minimal sample size of an observational study, it is sufficient under mild conditions to have two parameters additional to the standard inputs in the power calculation of randomised trials, which quantify the strength of the confounder-treatment and the confounder-outcome association, respectively. For the former, we propose using the Bhattacharyya coefficient, which measures the covariate overlap and, together with the treatment proportion, leads to a uniquely identifiable and easily computable propensity score distribution. For the latter, we propose a sensitivity parameter bounded by the R-squared statistic of the regression of the outcome on covariates. Utilising the Lyapunov Central Limit Theorem on the linear combination of covariates, our procedure does not require distributional assumptions on the multivariate covariates. We develop an associated R package PSpower.

Keywords: causal inference; observational study; overlap; power; sample size

Semiparametric Mediation Analysis Using Single-Index Models

Yen-Tsung Huang

Institute of Statistical Science, Academia Sinica

ABSTRACT

Mediation analysis is increasingly used to investigate how an exposure Z influences an outcome Y through a set of mediators $\boldsymbol{M}$. However, most existing methods either rely on parametric assumptions for the mediators and outcome, or are limited to binary exposures and a single mediator. We propose a novel algorithm that accommodates single-index models (SIMs) with non-binary exposures and adjustment for potential confounders, while avoiding parametric assumptions for both mediators and outcome. Specifically, we introduce two SIMs: one modeling the outcome conditional on the exposure, mediators, and confounders, and the other modeling the mediators conditional on the exposure and confounders. We derive a mediation estimand of the form $\mathrm{E}\left[\mathrm{E}\left[Y_j|Z_j = z_a, \boldsymbol{M}_j = \boldsymbol{M}_i\right]|Z_i = z_b\right]$, representing the expected outcome under a joint intervention setting the exposure to z_a and the mediators to their counterfactual values under z_b. The proposed estimator proceeds in four steps: (1) regress Y on Z and $\boldsymbol{M}$ using an SIM; (2) predict Y given z_a and $\boldsymbol{M}$; (3) regress the predicted Y on Z using another SIM; (4) predict the outcome at z_b. We establish the asymptotic properties of this estimator and evaluate its finite-sample performance through simulation studies. Finally, we demonstrate the practical utility of our method with two applications. The first examines the effect of socioeconomic status on body mass index mediated by DNA methylation of the *FASN* (fatty acid synthase) gene. The second investigates the effect of obesity on glycemic control through triglyceride, fasting glucose, and urine microalbumin.

Keywords: causal inference; mediation analyses; multiple mediators; single-index models

Causal Mediation Analysis: A Summary-Data Mendelian Randomization Approach

Shu-Chin Lin[1,2], Sheng-Hsuan Lin[3], Tian Ge[4,5,6], Chia-Yen Chen[7], Yen-Feng Lin[1,8,9]

[1]*Center for Neuropsychiatric Research, National Health Research Institutes, Miaoli, Taiwan*

[2]*Institute of Statistics and Data Science, National Taiwan University, Taipei, Taiwan*

[3]*Institute of Statistics, National Yang Ming Chiao Tung University, Hsinchu, Taiwan*

[4]*Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA*

[5]*Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA*

[6]*Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA*

[7]*Translational Medicine, Biogen, Cambridge, Massachusetts, USA*

[8]*Department of Public Health & Medical Humanities, School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan*

[9]*Institute of Behavioral Medicine, College of Medicine, National Cheng Kung University, Tainan, Taiwan*

ABSTRACT

Summary-data Mendelian randomization (MR) has become a popular tool for causal mediation analysis, offering an alternative to traditional methods. Two MR-based approaches corresponding to the difference and product methods have been implemented using inverse-variance weighted estimation (MR-IVW). However, existing methods often lack statistical efficiency, robustness, and rigorous inference procedures. In this study, we develop improved MR-based mediation frameworks using summary-level data, typically from genome-wide association studies (GWAS) data. Our contributions are threefold: we propose new variance estimators for mediation effects, establish formal inference procedures, and develop robust strategies to account for pleiotropy.

Keywords: causal inference; indirect effect; mediation analysis; mediation proportion; summary-data Mendelian randomization.

Sobolev Gradient Ascent for Optimal Transport: Barycenter Optimization and Convergence Analysis

Kaheon Kim[1], Bohan Zhou[2], **Changbo Zhu[1]**, Xiaohui Chen[3], Arlina Shen

[1]*University of Notre Dame*

[2] University of California, Santa Barbara

[3] University of Southern California

ABSTRACT

This paper introduces a new constraint-free concave dual formulation for the Wasserstein barycenter. Tailoring the vanilla dual gradient ascent algorithm to the Sobolev geometry, we derive a scalable Sobolev gradient ascent (SGA) algorithm to compute the barycenter for input distributions supported on a regular grid. Despite the algorithmic simplicity, we provide a global convergence analysis that achieves the same rate as the classical subgradient descent methods for minimizing nonsmooth convex functions in the Euclidean space. A central feature of our SGA algorithm is that the computationally expensive c-concavity projection operator enforced on the Kantorovich dual potentials is unnecessary to guarantee convergence, leading to significant algorithmic and theoretical simplifications over all existing primal and dual methods for computing the exact barycenter. Our numerical experiments demonstrate the superior empirical performance of SGA over the existing optimal transport barycenter solvers.

Keywords: optimal transport; Wasserstein barycenter; concave dual; gradient ascent;

Integration of the Knowledge Space: Network Structure, Random Search and Synergy

Daeseong Kim[1], **Masato S. Abe[1,2,3]**

[1]*Faculty of Culture and Information Science, Doshisha University, Kyoto, Japan*

[2]*Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan*

[3]*CBS-TOYOTA Collaboration Center, RIKEN, Saitama, Japan*

ABSTRACT

One of the keys to humanity becoming one of the most prosperous species on Earth lies in our inherent tendency to expand knowledge—that is, to explore the unknown. For instance, the seemingly unceasing progress in technology and the continual stream of scientific discoveries can be attributed to our persistent drive to explore new phenomena. Although the process of discovering novel entities has not been extensively studied scientifically, theoretical approaches within complex systems science have sought to identify general principles underlying such exploration. In this presentation, I introduce a mathematical modeling approach that conceptualizes exploration in knowledge networks as a random walk process and shortest path discovery process, aiming to understand the patterns underlying novel discoveries. Furthermore, I will discuss integration of knowledge space at the collective level, examining how the composition of individuals within a group can affect the efficiency of the exploration process.

Keywords: Knowledge Network; Collective Intelligence; Synergy; Cultural Evolution

Building A Food Knowledge Graph: Integration of Food-related Data Sources

Naoki Yoshimaru[1], Kenji Hatano

Graduate School of Culture and Information Science, Doshisha University, Japan

ABSTRACT

As food-related information continues to grow in complexity and variety, there is an increasing need to organize and utilize it more effectively. Food plays a central role in human life, and food-related data including recipes, ingredients and nutrition, has been widely shared online in recent years. However, this data is often unstructured and scattered across different formats and media, making it difficult to utilize effectively across domains such as health, chemistry, and gastronomy.

To address this challenge, we propose a Food Knowledge Graph (FKG) that extracts and integrates heterogeneous food-related information from diverse sources including text, images, videos, and chemical data. By linking entities such as ingredients, nutrients, cooking methods, tastes, and health effects, the FKG provides a unified, graph-based structure that captures complex, cross-domain relationships.

Unlike traditional single-modality approaches, our method enables flexible, multimodal integration without being limited by data formats. In this talk, we present our framework for constructing the FKG and showcase applications that highlight its potential in food computing and related fields. We aim to make these technologies accessible and useful not only for specialists, but also for a wider range of food-related applications.

Keywords: Food Computing; Knowledge Graph; Heterogeneous data.

Reproducibility-Optimized Component Clustering: A Test-Retest Framework for Robust Intrinsic Network Identification

Arthur C. Tsai

Institute of Statistical Science, Academia Sinica, Taiwan

Email: *arthur@stat.sinica.edu.tw*

ABSTRACT

Reproducibility is a central challenge in unsupervised component analysis, where results may vary substantially across repeated decompositions. This work presents a data-driven statistical framework for reproducibility-optimized component clustering that objectively determines reliability thresholds using test–retest information. Building on the RAICAR (Ranking and Averaging Independent Component Analysis by Reproducibility) framework, we model component-level reproducibility scores as arising from a latent mixture of reliable and unreliable populations using a mixed multinomial formulation. Parameters of the mixture model are estimated from repeated decompositions, and an optimal reproducibility threshold is selected via receiver operating characteristic (ROC) analysis to maximize discrimination between the two latent classes. This approach replaces ad hoc or heuristic thresholding with a principled, decision-theoretic criterion. The resulting framework yields stable and interpretable component clusters across independent runs and provides a reproducible foundation for downstream inference, including connectivity analysis and group comparisons. Although motivated by EEG and fMRI applications, the methodology is broadly applicable to any modality involving independent component analysis or related unsupervised decompositions.

Keywords: Reproducibility; Test–retest reliability; Independent component analysis (ICA); Component clustering; Intrinsic connectivity networks (ICNs); RAICAR; ROC analysis; Mixed multinomial modeling; EEG; fMRI

Function-on-Function Prediction via Deep Generative Models

Tso-Jung Yen[1], Chia-Tse Wang[1], Ming-Chung Chang[1], Su-Yun Huang[1], Tailen Hsing[2]

[1]*Institute of Statistical Science, Academia Sinica, Taiwan*

[2]*Department of Statistics, University of Michigan, Ann Arbor*

ABSTRACT

Function-on-function prediction involves using one sequence to predict another sequence. It is a problem commonly seen in many scientific fields. Its model is trained on paired sequence data. However, sequences may have different lengths, observed at irregular spaces and at different locations. These irregularities make model training difficult to proceed. In this talk, we present a graph-based method for training deep generative models to tackle function-on-function prediction problems. This method addresses a common challenge in training data, where functional sequences are only partially observed at irregular locations. Under this method, a functional sequence is represented as a graph in which each node corresponds to a location–value pair, and each link is defined in terms of the distance between locations of two nodes. This formulation allows training data to have functional sequences with different lengths and observed at different locations. Simulation results show that deep generative models trained under our method outperform the ground-truth model when only incomplete observations are available.

Keywords: Denoising Diffusion Probabilistic Models; Functional data; Graph neural networks.

Novel Empirical Likelihood Method for the Cumulative Hazard Ratio under Stratified Cox Models

Yichuan Zhao[1], Dazhi Zhao[2,*]

Department of Mathematics and Statistics, Georgia State University

ABSTRACT

Evaluating the treatment effect is a crucial topic in clinical studies. Nowadays, the ratio of cumulative hazards is often applied to accomplish this task, especially when those hazards may be nonproportional. The stratified Cox proportional hazards model, as an important extension of the classical Cox model, has the ability to flexibly handle nonproportional hazards. In this article, we propose a novel empirical likelihood method to construct the confidence interval for cumulative hazard ratio under the stratified Cox model. The large sample properties of the proposed profile empirical likelihood ratio statistic are investigated, and the finite sample properties of the empirical likelihood-based estimators under some different situations are explored in simulation studies. The proposed method was finally applied to perform statistical analysis on a real-world dataset on the survival experience of patients with heart failure.

Keywords: Cumulative hazard ratio; Empirical likelihood; Nonproportionality; Stratified Cox model; Survival analysis

GUEST: Graphical Models for Ultrahigh-Dimensional and Error-Prone Data by the Boosting Algorithm

Li-Pang Chen, Hui-Shan Tsao

Department of Statistics, National Chengchi University

ABSTRACT

In bioinformatics studies, understanding the network structure of gene expression variables is one of the main interests. In the framework of data science, graphical models have been widely used to characterize the dependence structure among multivariate random variables. However, the gene expression data possibly suffer from ultrahigh-dimensionality and measurement error, which make the detection of network structure challenging. The other important application of gene expression variables is to provide information to classify subjects into various tumors or diseases. In supervised learning, while linear discriminant analysis is a commonly used approach, the conventional implementation is limited in precisely measured variables and computation of their inverse covariance matrix, which is known as the precision matrix. To tackle those challenges, we introduce a new method called GUEST, which refers to Graphical models for Ultrahigh-dimensional and Error-prone data by the booSTing algorithm. The estimation strategy includes measurement error correction in high-dimensional variables under various distributions and then applies the boosting algorithm to identify the network structure and estimate the precision matrix. When the precision matrix is estimated, it can be used to construct the linear discriminant function and improve the accuracy of the classification.

Keywords: Feature screening; measurement error; network structure; statistical learning

A Likelihood Approach for Data Integration Involving Missing Data and Misclassified Variables

Zheng Yu[1], **Hua Shen[2*]**

Department of Mathematics and Statistics, University of Calgary

ABSTRACT

Integrating information from probability and nonprobability samples for unbiased inference can encounter both missing data and measurement errors, distorting population estimates. We develop a likelihood-based method that addresses missing outcomes and misclassified covariates in a probability sample, as well as dual misclassification of both outcome and covariate in a nonprobability sample. Simulation studies show substantial bias reduction over naive methods, and a real-world example underscores the method's practical advantages for more reliable inference.

Keywords: Data Integration, Measurement Error, Nonprobability Sample

Trans-Ancestry Cell-Type-Specific eQTLs Mapping by Integrating scRNA-seq and Bulk Data

Wenxin Jiang[1], **Mingxuan Cai[1,*]**

[1]*Department of Biostatistics, City University of Hong Kong*

ABSTRACT

Genome-wide association studies (GWAS) have successfully identified numerous genetic variants associated with complex traits and diseases, primarily located in noncoding regions. The emergence of expression quantitative trait loci (eQTLs) studies offered a unique opportunity to connect these variations to gene expression in relevant biological conditions. Despite the promise, two major challenges remain for existing eQTL studies. First, many eQTL effects are cell-type-specific. Traditional eQTL analyses often rely on bulk RNA sequencing (bulk RNA-seq) data, which measures the average gene expression levels across heterogeneous cell types and states, obscuring the cell-type-specific genetic effects. While the recent advancements in single-cell RNA sequencing (scRNA-seq) allow for deeper investigation at the cell-type level, they are limited by increased technical noise and smaller sample sizes. Second, current eQTL findings are predominantly based on European samples, making it difficult to extrapolate these results to non-European ancestries and hindering the interpretation of GWAS findings in diverse populations.

Here, we introduce a unified framework for trans-ancestry cell-type-specific eQTLs (ct-eQTLs) mapping by integrating summary statistics from bulk RNA-seq and scRNA-seq datasets (traceCB). This approach not only controls type I errors at nominal levels but also enhances the statistical power for identifying ct-eQTLs in underrepresented populations. TraceCB leverages several unique features. First, it boosts the statistical power by leveraging a larger scRNA-seq data from the European population while accounting for ancestral heterogeneities. Second, using the scRNA-seq data as a bridge, it further improves ct-eQTL mapping by integrating a large bulk RNA-seq data (e.g., the GTEx cohort). Third, unlike existing meta-analysis methods, it effectively accounts for the heterogeneous eQTL effects across populations and cell types, yielding well-calibrated p-values. Fourth, it only requires summary statistics of eQTL studies as its input, making it widely applicable to various tissues, cell types, and populations.

We demonstrate the effectiveness of traceCB through extensive simulations and analyses of real datasets, including multiple sc-eQTL datasets from peripheral blood mononuclear cells and bulk eQTL data from the GTEx and eQTLGen consortia. Our results show that traceCB achieves a substantial power gain in ct-eQTL mapping compared to existing methods.

Colocalization analysis of traceCB output and GWAS data reveals novel cell-type-specific regulatory mechanisms, elucidating the genetic basis of complex traits in the African and East Asian populations at a cellular resolution.

Keywords: single-cell eQTL; Statistical Genetics; GWAS

Back to Sessions List

Online Stochastic Optimization with Offline Data

Jize Xie, **Yi Chen**, Rachel Q. Zhang

Department of Industrial Engineering and Decision Analytics

Hong Kong University of Science and Technology

ABSTRACT

For online decision-making under uncertainty, data arrives in a streaming manner and decisions must be made sequentially. As a prominent modeling tool, online stochastic optimization finds broad applications, with the online stochastic gradient descent (OSGD) algorithm being a standard solution. While much existing research focuses on the online process solely, offline data are often available ahead, offering opportunities to enhance the online decision-making's quality. In this work, we investigate this problem by integrating offline data into the standard OSGD algorithm. We first propose an algorithm called OSGDwO that leverages offline information for a better online process initialization and stochastic gradient variance reduction. When offline and online data are identically distributed, by introducing a novel analytical framework, we demonstrate that the OSGDwO algorithm attains regrets decaying at the rate of O(1/N), where N is the offline sample size. To mitigate the risk due to a potential shift between offline and online distributions, we further develop a robustified algorithm by incorporating sequential hypothesis tests. This algorithm adaptively determines whether to utilize offline data, achieving the near-optimal performance guarantee regardless of the distributional shift's magnitude. Extensive simulation studies demonstrate the improved and robust performances of our algorithms. Finally, we apply our algorithms to two operations management problems: auction bidding and inventory management, illustrating both the practicality and extensibility of our approaches.

Keywords: stochastic optimization, sequential analysis, regret analysis

Adaptive Bayesian Optimization with Consistent Smoothness Estimation and Hyperparameters Exploration

Youxin Wang[1], **Saifei Sun[2]**, Shouri Hu[3]

[1,3]*School of Mathematical Sciences, University of Electronic Science and Technology of China.*

[2]*Department of Biostatistics, City University of Hong Kong.*

ABSTRACT

We present a novel algorithm, Adaptive Matérn Kernel Bayesian Optimization (AMKBO), designed to address the challenges of hyperparameter uncertainty in Gaussian Process (GP)-based Bayesian Optimization. The proposed method accelerates convergence by consistently estimating the smoothness parameter of the Matérn covariance kernel, adjusting the length-scale hyperparameter adaptively, and incorporating an opposition-based learning mechanism into acquisition function optimization. Specifically, AMKBO generates initial sampling points using random curves and constructs an estimator for the smoothness parameter of the covariance kernel, allowing the kernel to model functions with varying degrees of smoothness. During optimization, a refined length-scale adjustment strategy is employed to achieve a more balanced trade-off between exploration and exploitation, while preventing excessively aggressive exploration. In parallel, the opposition-based learning mechanism enhances the search coverage and computational efficiency in acquisition function optimization. Experimental results on synthetic benchmark functions and real-world problems demonstrate that AMKBO outperforms existing methods by achieving faster convergence and avoiding local optima compared to existing methods.

Keywords: Gaussian processes, Bayesian Optimization, hyperparameter adaptation, Matérn covariance kernel, smoothness estimation, opposition-based learning.

A Preferential Latent Space Model for Text Networks

Maoyu Zhang[1], **Biao Cai[2],** Dong Li[3], Xiaoyue Niu[4], Jingfei Zhang[1]

[1]*Goizueta Business School, Emory University, Atlanta, GA*

[2]*Department of Decision Analytics and Operations, City University of Hong Kong, Hong Kong*

[3]*Department of Statistics and Data Science, Tsinghua University, Beijing, China*

[4]*Department of Statistics, Pennsylvania State University, University Park, PA*

ABSTRACT

Network data enriched with textual information, referred to as text networks, arise in a wide range of applications, including email communications, scientific collaborations, and legal contracts. In such settings, both the structure of interactions (i.e., who connects with whom) and their content (i.e., what is communicated) are useful for understanding network relations. Traditional network analyses often focus only on the structure of the network and discard the rich textual information, resulting in an incomplete or inaccurate view of interactions. In this paper, we introduce a new modeling approach that incorporates texts into the analysis of networks using topic-aware text embedding, representing the text network as a generalized multi-layer network where each layer corresponds to a topic extracted from the data. We develop a new and flexible latent space network model that captures how node-topic preferences directly modulate edge formation, and establish identifiability conditions for the proposed model. We tackle model estimation with a projected gradient descent algorithm, and further discuss its theoretical properties. The efficacy of our proposed method is demonstrated through simulations and an analysis of an email network.

Keywords: text networks; latent space model; non-convex optimization; text analysis; topic-aware embedding.

Penalized Generative Variable Selection

Shuangge Ma

Department of Biostatistics, Yale School of Public Health

ABSTRACT

Deep networks are increasingly applied to a wide variety of data, including data with high-dimensional predictors. In such analysis, variable selection can be needed along with estimation/model building. Many of the existing deep network studies that incorporate variable selection have been limited to methodological and numerical developments. In this study, we consider modeling/estimation using the conditional Wasserstein Generative Adversarial networks. Group Lasso penalization is applied for variable selection, which can improve model estimation/prediction, interpretability, and stability. Significantly advancing from the existing literature, the analysis of censored survival data is also considered. We establish the convergence rate for variable selection while considering the approximation error, and obtain a more efficient distribution estimation. Simulations and the analysis of real experimental data demonstrate satisfactory practical utility of the proposed analysis.

Keywords: Deep neural network; Variable selection; Penalization; Consistency

Sparse Representation Learning for Scalable Single-Cell RNA Sequencing Data Analysis

Kai Zhao[1], Hon-Cheong So[2], **Zhixiang Lin[1]**

[1]*Department of Statistics, The Chinese University of Hong Kong, Hong Kong SAR, China*

[2]*School of Biomedical Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China*

ABSTRACT

The rapid rise in the availability and scale of scRNA-seq data needs scalable methods for integrative analysis. Though many methods for data integration have been developed, few focus on understanding the heterogeneous effects of biological conditions across different cell populations in integrative analysis. Our proposed scalable approach, scParser, models the heterogeneous effects from biological conditions, which unveils the key mechanisms by which gene expression contributes to phenotypes. Notably, the extended scParser pinpoints biological processes in cell subpopulations that contribute to disease pathogenesis. scParser achieves favorable performance in cell clustering compared to state-of-the-art methods and has a broad and diverse applicability.

Keywords: Genomics; Integrative analysis; matrix factorization

Bin Nan

Bayesian Bi-directional Self-Exciting Threshold Autoregressive Models for Loss Reserving

Yuning Zhang[1], **Wilson Ye Chen[1]**, S.T. Boris Choy[1], Tak Kuen Siu[2]

[1]University of Sydney, [2]Macquarie University

ABSTRACT

This work introduces the BiSETAR model, a novel approach to loss reserving that incorporates threshold uncertainty within a coherent Bayesian framework. Stochastic models with time-evolving parameters are generally considered state-of-the-art in addressing challenges posed by sudden large claims and irregular claim behaviours, but they treat threshold parameters as hyperparameters and neglect the uncertainty associated with these thresholds. The key contribution of this work is the introduction of a Bayesian method that treats threshold parameters as model parameters, allowing for the direct quantification of their uncertainty through the posterior distribution. To implement this, an efficient MCMC sampling algorithm is developed to approximate the posterior distribution of the threshold parameters. By incorporating bi-directional dynamics and threshold uncertainty into the forecasting process, the model significantly enhances the robustness and accuracy of loss forecasts, providing a more coherent and reliable framework for loss reserving.

Keywords: Bayesian Modelling, Threshold Autoregressive Model, Loss Reserving, Run-off Triangle

Loss-based Bayesian Sequential Prediction of Value-at-Risk with a Long-Memory and Non-linear Realized Volatility Model

Rangika Peiris, Minh-Ngoc Tran, **Chao Wang**, Richard Gerlach*

Discipline of Business Analytics, The University of Sydney

ABSTRACT

A long-memory and non-linear realized volatility model class is proposed for direct Value-at-Risk (VaR) forecasting. This model, referred to as RNN-HAR, extends the heterogeneous autoregressive (HAR) model, a framework known for efficiently capturing long memory in realized measures, by integrating a Recurrent Neural Network (RNN) to handle the non-linear dynamics. Quantile loss-based generalized Bayesian inference with Sequential Monte Carlo is employed for model estimation and sequential prediction in RNN-HAR. The empirical analysis is conducted using daily closing prices and realized measures with around 12 years of data till 2022, covering 31 market indices. The proposed model's one-step-ahead VaR forecasting performance is compared against a basic HAR model and its extensions. The results demonstrate that the proposed RNN-HAR model consistently outperforms all other models considered in the study. The implementation code of the HAR-RNN model is publicly available on Github: https://github.com/chaowang-usyd/RNN-HAR.

Keywords: HAR model; Recurrent Neural Network; Quantile Score; Sequential Monte Carlo; Generalized Bayesian inference.

Estimating Heterogeneous Treatment Effects through Multilevel Modeling

Tu Duong Quach[1], **Nuttanan Wichitaksorn[1]**, Kiet Nguyen[2], and Joshua D. Hawley[3]

[1]*Department of Mathematical Sciences, Auckland University of Technology, New Zealand*

[2]*University of Economics Ho Chi Minh City, Vietnam*

[3]*John Glenn College of Public Affairs and Center for Human Resource Research, Ohio State University, USA*

ABSTRACT

We propose a novel design for event studies using multilevel/hierarchical modeling. We show that the average treatment effect on the treated can be recovered from the variance of the nested random intercepts, and the heterogeneous treatment effects can be decomposed into two components: homogeneous treatment effects and the difference in outcomes between treated units. To illustrate our methodology, we revisit two studies on minimum wage policy and hospitalization. We thereby point out the pitfalls of the pre-trend testing when comparing one or a nearby period and fail to capture the heterogeneity.

Keywords: Multilevel modeling, Hierarchical modeling, Difference-in-differences, Heterogeneous treatment effects

Predicting the Weekly Return Direction of the S&P 500 Index Using DNN for Time Series Classification

Sang-Hyeok Lee

Department of Big Data, Small Enterprise and Market Service

ABSTRACT

This study explores the short-horizon predictability of the S&P 500 index by reformulating return forecasting as a binary time-series classification problem. Using deep neural networks, we predict whether the average return of the upcoming week will exceed that of the current week. Among 48 candidate architectures, a multilayer perceptron (MLP) model achieved the highest test accuracy (71.62%), outperforming both convolutional alternatives and six classical machine learning models. Beyond statistical metrics, we assess the economic value of the model's predictions through four trading strategies: Buy & Hold, Buy & Sell, Accuracy-Weighted, and Precision-Weighted. The Buy & Sell strategy delivered a cumulative return of 9,151%, while the Precision-Weighted strategy achieved the highest Sharpe ratio (3.08) and lowest Value-at-Risk (VaR). These results demonstrate that directional signals derived from deep learning can yield economically meaningful and risk-adjusted excess returns in weekly equity forecasting. Our findings challenge the strong form of market efficiency by uncovering persistent short-term patterns in price dynamics. The results also underscore the value of incorporating model confidence—through accuracy and precision—into portfolio construction. This research contributes to the growing literature on deep learning in finance by offering a robust and interpretable framework for short-term prediction and strategy design.

Keywords: Deep Learning, Time Series Classification, S&P 500 Forecasting, Short-Term Predictability

Investigating Spatial Omics Data with StarTrail and STimage-1K4M

Yun Li[1,2], Jiawen Chen[2], Caiwei Xiong[2], Muqing Zhou[3], Quan Sun[2], Gaorav Gupta[4], Aritra Halder[5], Didong Li[2]

[1] *Department of Genetics, University of North Carolina, Chapel Hill*

[2] *Department of Biostatistics, University of North Carolina, Chapel Hill*

[3] *Curriculum of Bioinformatics and Computational Biology, University of North Carolina, Chapel Hill*

[4] *Department of Radiation Oncology, University of North Carolina, Chapel Hill*

[5] *Department of Epidemiology and Biostatistics, Drexel University*

ABSTRACT

Spatial omics technologies revolutionize studies of tissue functions. However, existing methods fail to capture localized, sharp changes characteristic of critical events such as tumor development. We present StarTrail, a gradient based method that powerfully defines rapidly changing regions and detects “cliff genes”, genes exhibiting drastic expression changes at highly localized or disjoint boundaries. StarTrail, filling important gaps in current literature, enables deeper insights into tissue spatial architecture. We also introduce STimage-1K4M, a comprehensive dataset designed to bridge this gap by providing transcriptomic features for sub-tile images. STimage-1K4M contains 1,149 images and 4,293,195 pairs of sub-tile images and gene expressions. STimage-1K4M offers unprecedented granularity, paving the way for a wide range of advanced research in multi-modal data analysis.

Keywords: spatial omics; gradient; Gaussian process; histological image

Multi-Ancestry Fine-Mapping of Causal Variants in Genome-Wide Association Studies

Boran Gao[1], **Xiang Zhou[2,*]**

[1]*Department of Statistics, Purdue University*

[2]*Department of Statistics and Data Science, Yale University*

ABSTRACT

Fine-mapping in genome-wide association studies attempts to identify causal SNPs from a set of candidate SNPs in a local genomic region of interest and is commonly performed in one genetic ancestry at a time. Here, we present multi-ancestry sum of the single effects model (MESuSiE), a probabilistic multi-ancestry fine-mapping method, to improve the accuracy and resolution of fine-mapping by leveraging association information across ancestries. MESuSiE uses summary statistics as input, accounts for the diverse linkage disequilibrium pattern observed in different ancestries, explicitly models both shared and ancestry-specific causal SNPs, and relies on a variational inference algorithm for scalable computation. We evaluated the performance of MESuSiE through comprehensive simulations and multi-ancestry fine-mapping of four lipid traits with both European and African samples. In the real data, MESuSiE improves fine-mapping resolution by 19.0% to 72.0% compared to existing approaches, is an order of magnitude faster, and captures and categorizes shared and ancestry-specific causal signals with enhanced functional enrichment.

Keywords: GWAS; fine-mapping; multi-ancestry; Bayesian

An Alternative Method for Instrument Variable Regression: Reverse Two-Stage Least Squares (r2SLS)

Lei Fang, **Wei Pan**

Division of Biostatistics and Health Data Science, University of Minnesota

Email: panxx014@umn.edu

ABSTRACT

Two-stage least squares (2SLS) is by default applied to infer a putative causal association between an exposure, such as a gene or a protein, with an outcome such as a complex disease or trait, in transcriptome- or proteome-wide association studies (TWAS/PWAS). In a typical two-sample setting for TWAS/PWAS, the stage 1 sample size is much smaller than that of stage 2. To reduce the resulting attenuation bias and estimation uncertainty in stage 1 and boost statistical power of the conventional TWAS, we propose a new method, called reverse two-stage least squares (r2SLS): instead of imputing a gene's expression (using genetic variants as instrumental variables, IVs) in stage 1 and then testing the association between the imputed expression and the observed outcome in stage 2 in the conventional 2SLS approach, we propose predicting the outcome (using IVs) and testing the association between the predicted outcome and the observed gene expression. Theoretically, we establish that the r2SLS estimator is asymptotically unbiased with a normal distribution. We also show theoretically when 2SLS and r2SLS are asymptotically equivalent and when r2SLS is asymptotically more efficient than 2SLS. We also consider the practical issue of how to select invalid IVs. We use simulations and three real data examples based on the GTEx gene expression data, UKB-PPP proteomic data and several GWAS summary datasets to demonstrate some advantages of r2SLS over 2SLS, including possibly better type I error control, higher statistical power and robustness to weak IVs.

Keywords: 2SLS; Causal inference; GWAS; TWAS

Partitioned Blood Pressure Polygenic Risk Reveals Differential Genetic Effects and Environmental Modulation of Cardiovascular Disease

Xiaofeng Zhu[1], Yihe Yang[1], Dongwon Lee[2], Aravinda Chakravarti[3]

[1]*Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, OH, 44106, USA*

[2]*Department of Pediatrics, Division of Nephrology, Boston Children's Hospital, Boston, MA, USA.*

[3]*Center for Human Genetics and Genomics, NYU Grossman School of Medicine, New York, NY, 10016, USA*

ABSTRACT

Genetic variants within cis-regulatory elements (CREs) contribute substantially to the heritability of complex traits, with tissue-specific effects shaping their regulatory functions. Constructing polygenic risk scores (PRSs) based on tissue-restricted CREs can provide deeper insights into the biology underlying complex traits. In this talk, we present a method to partition blood pressure (BP) PRSs according to CREs from four tissue types: adrenal gland, artery, heart, and kidney. These CRE-partitioned PRSs exhibit heterogeneous contributions to cardiovascular disease, reveal interactions with environmental factors, and improve predictive performance.

Keywords: polygenic risk scores, cis-regulatory elements, blood pressure, cardiovascular disease

Improving Unbiasedness of the Proportional Hazards Model Estimator with Cox and Snell's Bias Approximation and Jackknife Resampling

Chia-Hui Huang[1], Ruby Chiu-Hsing Weng[1]

Department of Statistics, National Chengchi University, Taipei, Taiwan

ABSTRACT

Bias approximation has played an important role in maximum likelihood estimation method, and numerous bias calculation techniques have been proposed under different contexts. For the semiparametric proportional hazards model, which is the standard regression method to study the time-to-event data, the existing work applied the bias formula and derived the approximate bias of Cox's estimator based on the partial likelihood function. In this work, we instead use the joint likelihood function and utilize the counting process approach to develop approximate bias of Cox's estimator. Explicit expressions for the higher order partial derivatives are derived, which facilitate the bias calculation techniques. We also incorporate jackknife resampling method and propose a Jackknife-Cox-Snell method that processes the biasedness of Cox's estimator through two steps, where the first step aims for removing the analytical terms derived from Cox and Snell's formula and the second step reduces the residual bias term then. A comprehensive simulation study is performed to show the usefulness of the proposed bias-corrected method. We also apply the proposed method to two sets of survival data for comparison and illustration.

Keywords: Bias calculation, Counting process, Cox and Snell's formula, Jackknife, Nonparametric maximum likelihood estimator, Proportional hazards model

Prediction-Oriented Transfer Learning for Semiparametric Transformation Models with Survival Data

Yu Gu[1,*], Donglin Zeng[2], D. Y. Lin[3]

[1]*Department of Statistics and Actuarial Science, The University of Hong Kong*

[2]*Department of Biostatistics, University of Michigan*

[3]*Department of Biostatistics, University of North Carolina at Chapel Hill*

ABSTRACT

Transfer learning is beneficial for survival analysis, especially when the target study has a limited number of events. However, existing transfer learning methods rely on restrictive assumptions that the target and source studies share similar parameters under Cox models. In this talk, I will introduce a novel transfer learning framework that enhances survival prediction by transferring predictive rather than distributional knowledge from source studies. Our approach employs flexible semi- parametric transformation models for the target data and eliminates the need to model or share the source data. The ingeniously designed penalty enables optimization via a simple and stable EM algorithm. We rigorously establish the asymptotic properties of the proposed estimator and show that it achieves a faster convergence rate than the target-only estimator when source knowledge is sufficiently accurate. We demonstrate the effectiveness of our methods through extensive simulation studies and an application to two major breast cancer studies.

Keywords: EM algorithm, Nonparametric likelihood, Survival prediction, Transfer learning, Transformation models

A Novel Approach When Facing Emerging Infectious Disease --- Randomized Selection Design

Cheng-Shiun Leu, Ph.D.

Columbia University Irving Medical Center, USA

ABSTRACT

When facing an emerging infectious disease (such as COVID-19 or Ebola) with no readily available effective treatment, one is less interested in making declarations of statistical significance than in identifying a promising direction forward. A randomized selection design offers the appropriate tools for the approach. The general goal of a randomized selection design is to select one or more treatments from several competing candidates to which patients are randomly assigned, in such a way that selected treatment(s) are likely to be better than those not selected. For example, if one treatment is clearly superior to all the others, we may demand that the procedure select that treatment with high probability. The experimental treatments could be different doses of a drug or intensities of a behavioral intervention, different treatment schedules, modalities, or strategies, or different combinations of treatments. The hallmark feature of a selection design is its ability to achieve its stated goals with surprisingly fewer participants compared with traditional "phase III" trials, precisely because it eschews the formal hypothesis test paradigm with its tight control over type I error rates. In addition, the selection procedure allows for sequential elimination of inferior treatments, sequential recruitment of superior treatments, and may be used to select treatments with fixed or variable subset sizes make it even more appealing in practice. Thus, they are ideal for middle development settings where we are interested in selecting promising treatments under circumstances typically limited by smaller sample sizes. In this talk, I'll focus on the discussion of the Levin-Robbins-Leu (LRL) family of sequential subset selection procedures, their design considerations, and actual applications in clinical trials.

Poster

No.	Name	Topic
1	Wei-Chun Kang	Cyclic SOAs and Moving Window Criteria for Space-Filling Designs
2	Li-Sheng Zhuang	Nonparametric Mediation Analysis of Non-Markov Illness-Death Model
3	Aman Prakash	Order Restricted and Unrestricted Inference for the Recursive Stage Life Testing Model
4	Anil Maurya	Analysing Load-Sharing System under Progressive Censoring Using Statistical and Machine Learning Approach
5	Vaibhav N. Dhameliya	Compound Optimal Design Strategy for Life-Testing Experiment under Progressive Hybrid Censoring Scheme
6	Chien-Tai Lin	Optimal Progressively Censored Reliability Sampling Plans for the Log-Location-Scale Distribution
7	Erjia Cui	Quantifying Physical Activity Intervention Effects via Functional Regression
8	Yun-Ting Wei	Comparison and Ensemble Strategies of Measurement Error Correction Methods in Multiple Regression with Validation Data
9	Jihoon Kim	A Deep Learning Random-Effect Modelling Approach for clustered Count Data
10	Chin-Sheng Teng	A Statistical Framework Leveraging Single-Cell Data for Cell Type Deconvolution in Bulk Transcriptomics
12	Sung-Hyuk Choi	Prediction of Deterioration of Patients with Dyspnea in Emergency Department
13	En-Yu Lai	Reframing Cross-World Independence for Identifying Path-Specific Effects
14	Shu-Hsien Cho	VACANT: An Adaptive Framework for Rare-Variant Association Testing that Leverages Continuous Functional Annotations
15	Chi Chun Yeh	Causal Mediation Analysis with Survival Data and a Recurrent Mediator
16	Kai-Yuan Wu	Authoritarian Regime Types and Their Pathways to Democratization via Coups
17	Mei Dong	Marginal Causal Effect Estimation with Continuous Instrumental Variables
18	Wei-Yang Yu	Automated Analysis of Experiments using Hierarchical Garrote
19	Jiuqian Shang	Uncertainty Quantification for Noisy Low-tubal-rank Tensor Completion
20	Elvin Tseng	Linearly Constrained Symmetric Rank-One Approximation for Pre-Image Recovery in Nonlinear Matrix Completion
21	Jia-Ying Su	Investigating the Pleiotropic Effects of Splice-Affecting Variants on Cancer and Metabolic Diseases
23	Mengqi Lin	Controlling the False Discovery Proportion in Observational Studies with Hidden Bias
24	Lars Skaaret-Lund	LBBNN: An R Package for Sparse Bayesian Neural Networks
25	Tianyu Liu	Cramér-Type Moderate Deviation and Berry–Esseen Bounds in the p-Spin Curie–Weiss Model
26	Kong Xin	Concentration-QTC Analysis to Support ICH E14 with a Real Case

Back to Sessions List

Cyclic SOAs and Moving Window Criteria for Space-Filling Designs

Wei-Chun, Kang, Cheng-Yu, Sun

National Tsing Hua University, Taiwan

ABSTRACT

Space-filling designs are essential in computer experiments to ensure that design points are uniformly distributed across the input space. Among them, strong orthogonal arrays (SOAs) are widely used for their stratification properties on fixed grids formed by collapsing adjacent factor levels. However, these grids lack flexibility and cannot adapt to local structures. A moving window approach is introduced that evaluates uniformity over sliding regions across the space. By specifying a window size and selecting a subset of positions, a more fine-grained space-filling criterion is defined that generalizes several existing methods. Within this framework, cyclic SOAs are further proposed: SOAs that preserve their stratification properties under cyclic shifts of factor levels. These designs exhibit a structural invariance that is particularly useful in settings involving periodicity or level relabeling. Their optimality properties are established, and construction methods are presented, positioning cyclic SOAs as a flexible and robust addition to the space-filling design toolkit.

Keywords: Computer experiments, Design of experiments, Orthogonal arrays, Strong orthogonal arrays, Centered L_2-discrepancy, Wrap-around L_2-discrepancy, Rectangular Uniformity criterion, Cyclic strong orthogonal arrays

Nonparametric Mediation Analysis of Non-Markov Illness-Death Model

Li-Sheng Zhuang[1], Jih-Chang Yu[2], Yen-Tsung Huang[3]

[1,3]*Institute of Statistical Science, Academia Sinica*

[2]*Department of Statistics, National Taipei University*

ABSTRACT

The illness-death model is widely used to characterize disease progression over time. Previous work focuses either on estimation without causal interpretation or on causal interpretation under a strong Markov assumption where the terminal event depends on the status but not the timing of the intermediate event. To bridge the research gap, we propose a new definition of counterfactual hazard that relaxes the Markov assumption by considering the entire history of the intermediate event. We derive an identification formula that involves an integral with respect to the probability density function of the intermediate event time. Direct and indirect effects refer to the influence of an exposure on the terminal event not mediated by, and mediated through, the intermediate event, respectively. We propose nonparametric kernel estimators for the two effects and study their asymptotic properties. We conduct numerical simulations to examine the proposed estimators' finite-sample performance. Applying the method to a hepatitis study where the Markov assumption is violated, we show that the effect of hepatitis C on mortality is not mediated through septicemia.

Keywords: illness-death model, non-Markov assumption, causal inference, mediation analysis, kernel density estimation

Order Restricted and Unrestricted Inference for the Recursive Stage Life Testing Model

Aman Prakash[1*], Raj Kamal Maurya[1], Debashis Samanta[2], Debasis Kundu[3]

[1]*Department of Mathematics, Sardar Vallabhbhai National Institute of Technology, Surat, India*

[2]*Department of Mathematics and Statistics, Aliah University, Kolkata, West Bengal, India*

[3]*Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, Kanpur, India*

ABSTRACT

In this study, we propose a recursive stage life testing (RSLT) framework that advances stage life testing experiments by allowing each surviving unit to move sequentially through multiple stages. In this setup, any unit that survives one stage moves on to the next, facing higher stress levels step by step. We develop classical inference procedures for the RSLT model under order-restricted and unrestricted scenarios. To obtain the estimate of the model parameters, maximum likelihood estimators have been used for both cases. Under unrestricted inference, parameters are estimated freely, and in order-restricted inference, we incorporate the natural assumption that higher stress levels lead to early failure. By comparing these two approaches, we see how using known patterns in the data can improve the accuracy of the estimates. We have discussed the proposed methodology using three widely used lifetime distributions: exponential, Weibull, and Chen. An extensive Monte Carlo simulation study evaluates estimator performance under unrestricted and order-restricted settings for each distribution. To illustrate its practical relevance, the proposed RSLT framework is applied to an aerospace electrical connector failure dataset, validating the flexibility and accuracy of the model.

Keywords: Stage life testing; Progressive censoring; Proportional hazard model; Maximum likelihood estimates; Order restriction; Confidence interval

Analysing Load-Sharing System under Progressive Censoring Using Statistical and Machine Learning Approach

Anil Maurya[1*], Raj Kamal Maurya[1]

[1]*Department of Mathematics, Sardar Vallabhbhai National Institute of Technology, Surat, India*

ABSTRACT

This study focuses on the reliability analysis of a multi-component load-sharing system under a progressive censoring scheme. In such a system, where a component fails, its load is redistributed among the surviving components, increasing their failure rates. Such systems are known as load-sharing systems. The lifetimes of the components follow the Weibull distribution; we estimate system parameters and reliability measures using both classical and Bayesian frameworks. In the classical approach, the maximum likelihood estimation method is employed, with asymptotic confidence intervals derived for system parameters. For the Bayesian framework, we use informative priors to obtain Bayesian estimates and construct the highest posterior density intervals. A Monte Carlo simulation study is presented to compare the performance of these estimation methods. Furthermore, real data is used to validate the proposed methodologies, demonstrating the effectiveness of the approaches. Furthermore, we introduce machine learning techniques to analyze the reliability of the model. We have employed three different machine learning methods, such as support vector regressor, random forest regressor, and gradient boosting regressor, to estimate reliability. A comparative analysis of all three methods has been discussed using performance metrics. Finally, the influence of model parameters on reliability estimation has been explored.

Keywords: Load-sharing system, Progressive censoring scheme, Maximum likelihood estimator, Bayes estimator, Machine learning prediction, Performance evaluation

Compound Optimal Design Strategy for Life-Testing Experiment under Progressive Hybrid Censoring Scheme

Vaibhav N. Dhameliya[1*], Raj Kamal Maurya[1], Ritwik Bhattacharya[2]

[1]*Department of Mathematics, Sardar Vallabhbhai National Institute of Technology, Surat, India*

[2]*Department of Mathematical Sciences, University of Texas at El Paso, El Paso, Texas, USA*

ABSTRACT

In life-testing experiments, optimal designs based on single objectives or constraints have been extensively studied, whereas research on multi-objective approaches is still in its early stages of development. In this study, we have introduced a compound design strategy for optimizing life-testing plans under a progressive hybrid censoring scheme. Our approach utilizes a graphical method to develop and evaluate efficient test designs, balancing cost with factors such as trace, determinant, and variance of the inverse Fisher information matrix. With the help of illustrative examples, we demonstrate the advantages of compound optimal designs compared with traditional constraint-based methods. Furthermore, we have discussed the advantages of compound optimal designs over traditional constraint-based methods and used sensitivity analysis to assess design robustness and limitations. Additionally, we analyze real-life data to validate the practical applicability of our proposed methodology, showcasing its effectiveness in optimizing life-testing experiments.

Keywords: Life testing, Compound optimal design, Weibull distribution, Cost function, Progressive hybrid censoring scheme, Sensitivity analysis

Optimal Progressively Censored Reliability Sampling Plans for the Log-Location-Scale Distribution

Chien-Tai Lin[1,*], Shao-Lun Jane[1] and N. Balakrishnan[2]

Department of Mathematics, Tamkang University, Taiwan

Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada

ABSTRACT

We introduce here a variable neighborhood search (VNS) algorithm-based approach to determine the minimum sample sizes required for progressively censored reliability sampling plans within the flexible family of log-location-scale family of distributions, which includes Weibull and log-logistic distributions. The proposed method significantly reduces sample sizes compared to previously developed approaches, demonstrating its feasibility, especially for small sample sizes, in contrast to complete search methods (CSM). Optimal censoring plans are identified using A- and D-optimality criteria, and a variance-measure criterion. The proposed approach consistently outperforms the method discussed in Ng et al. (2004) for the case of Weibull distribution.

Keywords: Acceptance sampling; Log-logistic distribution; Optimal criteria; Variable neighborhood search algorithm; Weibull distribution

Quantifying Physical Activity Intervention Effects via Functional Regression

Nidhi Pai, Yu Lu, Kristin A. Linn, **Erjia Cui**

Division of Biostatistics and Health Data Science, University of Minnesota, Minneapolis, Minnesota, U.S.A.

ABSTRACT

Physical activity (PA) intervention studies often collect repeated intensity measurements over long observation periods. Quantifying the variation in intervention effects over the study period is critical to evaluating and improving intervention strategies, yet many analyses reduce PA data into scalar summary measures, resulting in limited insights. We propose a functional regression framework, which captures time-varying intervention effects by modeling the entire PA trajectory as a functional observation. From both methodological and practical perspectives, we demonstrate the advantages of function-on-scalar regression (FoSR) over the traditional two-step approach of applying functional principal components analysis (FPCA) followed by regressing scores on covariates. The FoSR is further extended to a function-on-function regression (FoFR) for studying the association of PA across time periods. Methods are applied to daily step counts from the Social incentives to Encourage Physical Activity and Understand Predictors (STEP UP) study, revealing distinct and highly interpretable time-varying effects of three intervention strategies on PA and differences in their sustainability. Our case study highlights the feasibility of functional data analysis techniques for uncovering novel insights in intervention studies with high-dimensional endpoints.

Keywords: Functional data analysis; Functional regression; Physical activity intervention; Wearable devices

Comparison and Ensemble Strategies of Measurement Error Correction Methods in Multiple Regression with Validation Data

Yun-Ting Wei, Jia-Ren Tsai*

Department of Statistics and Information Science, Fu Jen Catholic University, New Taipei City, Taiwan (ROC)

ABSTRACT

Validation data provide researchers with access to replicate or gold-standard measurements on a subset of observations, which in turn facilitate the direct estimation of error variances within measurement error models. In the absence of validation data, measurement error is typically non-identifiable, leading to systematic attenuation bias in the estimation of regression coefficients. However, by incorporating validation data, the observed variance can be decomposed into true variance and error variance, allowing for the estimation of reliability ratios or calibration factors. These quantities are subsequently incorporated into correction procedures, yielding regression coefficients and variable selection outcomes that more closely approximate those based on the unobserved true values. This study integrates several methods with measurement error correction capabilities and further proposes an ensemble estimation strategy to leverage the strengths of different models, aiming to achieve optimal predictive performance. The simulation study encompasses both structural and functional models, settings with independent and correlated covariates, varying sample sizes, and different levels of measurement error. We examine how changes in the proportion of error-free observations within the validation sample affect model performance, thereby providing a comprehensive assessment of the applicability and predictive accuracy of each method and ensemble strategy across diverse scenarios. The results indicate that the presence of validation samples substantially reduces estimation bias and improves predictive accuracy. As the proportion of error-free observations increases, differences across methods gradually diminish, and their overall performance converges. Further comparisons under varying conditions reveal that the proposed ensemble strategy consistently achieves the lowest mean squared prediction error across all simulation scenarios as well as in empirical analyses of Visceral Adipose Tissue data, demonstrating superior robustness and accuracy in diverse settings. These findings underscore the critical role of validation data design and confirm the practical value of ensemble strategies for empirical research involving measurement error.

(* corresponding author: 141628@mail.fju.edu.tw)

Keywords: Measurement error model, Validation data, Ensemble strategy

A Deep Learning Random-Effect Modelling Approach for clustered Count Data

Jihoon Kim[1], VU TUAN ANH[2], Il Do Ha[1]

[1]*Department of Statistics and Data Science, Pukyong National University, Busan, South Korea*

[2]*Department of Artificial Intelligence Convergence, Pukyong National University, Busan, South Korea*

ABSTRACT

The deep neural network (DNN) model, a core model of deep learning, provides high predictive power to predict and classify output variables involved in various fields by modeling the nonlinear functional relationship between input and output variables through hidden layers. However, since the DNN has been mainly developed when output variables are independent, applying it immediately when there is a correlation between output variables has a disadvantage in that predictive performance is greatly reduced. This correlation is usually modeled via random effects, but the random effect model has been mainly developed under the assumption of linearity in the functional relationship between input and output variables.

In this talk, we propose a new Poisson DNN random-effect model. The output variables are correlated count outcomes which are obtained from the repeated measures over time. For estimation (learning) of the proposed model, we develop an optimization algorithm based on negative marginal likelihood as a loss function. Simulation and real data analysis demonstrate the validity of the proposed method. In particular, the simulation results confirmed that the proposed DNN model provides higher predictive performance than the existing prediction models.

Keywords: DNN, random effect, count data, marginal likelihood

A Statistical Framework Leveraging Single-Cell Data for Cell Type Deconvolution in Bulk Transcriptomics

Chin-Sheng Teng[1], Pei-Ling Lin[1], Yijia Xue[1], Ying Lin[1], Ryan Kitagawa[1], Makena Grigsby[1], RJ Honicky[2], Eugene Bolotin[2], Xiaoqian Liu[1], Weixin Yao[1], Xinping Cui[1,*]

[1]*Department of Statistics, University of California, Riverside, USA*

[2]*Industry Partner, Miraomics, USA*

ABSTRACT

Bulk transcriptomics measures averaged gene expression across all cells in a sample, masking the contributions of individual cell types. Abnormal cell type proportions are closely linked to disease initiation and progression, making accurate deconvolution both biologically and clinically important. Single-cell RNA sequencing (scRNA-seq) provides detailed cellular resolution, but its use is limited by cost and sample availability. A framework is therefore needed to leverage existing scRNA-seq data to develop and evaluate deconvolution methods for bulk datasets.

We developed a framework using real scRNA-seq data to generate pseudo-bulk mixtures with known ground-truth proportions, ensuring that natural biological heterogeneity is preserved. Within this framework, we introduced an alternative statistical approach and compared it with conventional deconvolution methods. Conventional approaches rely on predefined reference profiles for each cell type and use regression to estimate cell type proportions. In contrast, our approach directly learns gene-level weights that map bulk expression to cell type composition, reframing deconvolution as a supervised learning problem. This design enables flexible statistical modeling, including regression-based methods such as ridge and lasso regression, and tree-based methods such as Random Forest and XGBoost.

We applied our framework to the CZ CellxGene hippocampus scRNA-seq dataset to generate pseudo-bulk mixtures for evaluation. Once trained, the same models can be applied to bulk RNA-seq data such as GTEx, where true cell type proportions are not available, to estimate cellular composition in real tissue samples. While the hippocampus dataset is relatively small and limits generalization, our framework is flexible and scalable, making it well suited for application to larger single-cell and bulk datasets. This provides a practical tool for advancing statistical methods for cell type deconvolution in bulk transcriptomics.

Keywords: Cell type deconvolution, Bulk transcriptomics, Single-cell RNA-seq, Statistical framework, Supervised learning

Prediction of Deterioration of Patients with Dyspnea in Emergency Department

Sung-Hyuk Choi[1], Byeong-Cheul Ko[2]

[1]*Korea University Guro Hospital*

[2]*Korea Medical Institute*

ABSTRACT

The initial severity triage of patients with dyspnea is essential to provide adequate care. The appropriate value of the factor to 154,383 patients visiting the emergency department from January 1, 2015 to December 31, 2018. Actual hospitalization and discharge in patients with dyspnea who are visiting the emergency department from January 1, 2019 to December 31, 2020. The rapid predicition of hospitalization and discharge using Rasch analysis in ED was highly accurate when combined with more efficient factors, similar to the analysis of artificial intelligence.

Keywords: triage, severity, dyspnea

Reframing Cross-World Independence for Identifying Path-Specific Effects

En-Yu Lai[1], Jih-Chang Yu[2], Yen-Tsung Huang[1]

[1]*Institute of Statistical Science, Academia Sinica*

[2]*Department of Statistics, National Taipei University*

ABSTRACT

The challenge of identifying causal mechanisms in real-world problems often involves multiple factors and requires the evaluation of path-specific effects within a multi-mediator setting. Identification in this context depends not only on standard causal assumptions but also on the demanding cross-world independence assumption. To address this issue, Lin et al. (2017) introduced an alternative causal framework using an interventional approach, which fulfills the cross-world independence by redefining path-specific effects. Later, Stensrud et al. (2021) proposed the dismissible component conditions assumption to identify separable effects in the presence of competing events. In this study, we employed SWIGs to systematically investigate the underlying causal concepts of the three causal semantics for identifying path-specific effects. Specifically, we extended the notion of separable effects and formulated the corresponding assumptions required for identifying path-specific effects. We emphasized that violations of cross-world independence arise from mediators excluded from the model. By analogy with how exchangeability between actual and counterfactual outcomes is achieved through sufficient control of confounders, we argue that cross-world independence can be approximated in practice by incorporating a sufficient set of mediators.

Keywords: cross-world independence, generalized separable effects, interventional approach, path-specific effect, causal assumptions, multimediator model, SWIG

VACANT: An Adaptive Framework for Rare-Variant Association Testing that Leverages Continuous Functional Annotations

Shu-Hsien Cho[1,2], Yao Yu[2], Ryan Bohlender[2], Chad Huff[2]

[1]*UTHealth Graduate School of Biomedical Sciences*

[2]*Department of Epidemiology, University of Texas MD Anderson Cancer Center*

ABSTRACT

To address this, we developed the Variant Annotation Clustering Association Test (VACANT), a novel and flexible statistical framework. VACANT first partitions rare variants (MAF < 0.5%) into ordered, disjoint sets based on a continuous annotation score, a strategy designed to isolate association signals by stratifying variants according to their functional impact. The framework offers two testing modalities: 1) a univariate approach, where each variant set is independently tested and the resulting p-values are combined using the Aggregated Cauchy Association Test (ACAT), effectively preserving the signal from high-impact variant set; and 2) a multivariate approach, where all sets are simultaneously included as predictors in a single model, with significance assessed via a comprehensive goodness-of-fit test. For parameter estimation in both modalities, VACANT employs Firth's penalized logistic regression to ensure robust and well-calibrated results, even with sparse variant counts. To account for population structure and genetic relatedness, the framework can optionally incorporate a genetic relatedness matrix (GRM), shifting the underlying statistical mechanism to a generalized linear mixed model (GLMM).

Through extensive simulations and analysis of whole-exome sequencing data for key breast cancer susceptibility genes (ATM, BRCA1/2, CHEK2, PALB2) from the UK Biobank, we demonstrate the utility of our approach. VACANT consistently achieved similar or higher empirical power than established methods, including SKAT-O and STAAR-O. For instance, our method surpassed 95% power for BRCA1 and BRCA2 with approximately 2,000 cases, a threshold at which competing methods showed negligible power. Critically, VACANT maintained rigorous control of the Type I error rate across all tested scenarios. Its computational efficiency and modular architecture make VACANT a powerful and statistically robust framework for enhancing rare-variant association discovery in modern biobank-scale analyses.

Keywords: Rare-Variant Association Test, Functional Annotation, Case-Control Imbalances, Biobank Data

Causal Mediation Analysis with Survival Data and a Composite Time-to-Event Mediator

Chi-Chun Yeh[1] and Yen-Tsung Huang[2]

[1,2]*Institute of Statistical Science, Academia Sinica*

ABSTRACT

Mediation analysis aims to investigate how an exposure affects an outcome through mediators. It decomposes the effect from the exposure to the outcome into the effect through mediator(s) (indirect effect) and the effect direct to the outcome (direct effect). Recently, mediation analysis has been proposed for the setting where both mediator (e.g., diseases) and outcome (e.g., death) are time-to-event variables. In medicine, many disease categories or syndromes encompass simpler component diseases, as seen in Cardiovascular Kidney Metabolic (CKM) Syndrome. In this study, we build on this concept by considering a composite mediator, in which each component is a time-to-event variable. We propose a counterfactual hazard difference as our effect estimand and derive a counting process-based estimator, which can be simplified as a Nelson-Aalen estimator with time-varying weights. Asymptotic properties, including consistency and asymptotic normality, of the estimator are established using the martingale theory. Extensive numerical experiments are conducted to evaluate the performance of our proposed estimator. We demonstrate the method by applying it to REVEAL data to assess how HCV causes death through multiple chronic diseases (i.e., the CKM Syndrome).

Keywords: Causal inference, Mediation analysis, Mediation group

Authoritarian Regime Types and Their Pathways to Democratization via Coups

Kai-Yuan Wu[1], Jih-Chang Yu[2], Wen-Chin Wu[3], Yen-Tsung Huang[4,*]

[1]*Institute of Statistical Science, Academia Sinica, Taipei*

[2]*Department of Statistics, National Taipei University*

[3]*Institute of Political Science, Academia Sinica, Taipei*

[4]*Institute of Statistical Science, Academia Sinica, Taipei*

ABSTRACT

The relationship between coups and democracy has long been debated. However, the mechanisms through which authoritarian regimes transition to democracy remain unclear. This study investigates how different types of authoritarian regimes transition to democracy, using a causal mediation framework within a semi-competing risks setting. It focuses particularly on the role of coups. We compiled data from 116 countries that experienced authoritarian rule between 1950 and 2020. In the proposed framework, regime type is treated as the exposure, time to coup as the mediator, and time to democratization as the outcome. The exposure and confounders are time- varying, allowing past country-specific information to improve estimates of the causal relationships between authoritarian regimes, coups, and democratic transitions. We employ a semiparametric estimator to identify two effects: an indirect effect, the effect of regime type on democratization via coups, and a direct effect, the effect not mediated through coups. Our proposed semiparametric mediation analyses show that the original type of authoritarian regime affects the chance of later democratization, which may be mediated through the occurrence of coups. Our findings provide new insights into the pathways through which authoritarian regimes either maintain power or transition to democracy, underscoring that the effect of coups depends on the type of authoritarian regime.

Keywords: causal mediation model; semi-competing risk; coups; democracy

Marginal Causal Effect Estimation with Continuous Instrumental Variables

Mei Dong[1], Lin Liu[2], Dingke Tang[3], Geoffrey Liu[4], Wei Xu[1,5], Linbo Wang[6]

[1]*Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto*

[2]*Institute of Natural Sciences, MOE–LSC; School of Mathematical Sciences, CMA–Shanghai; SJTU–Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University*

[3]*Department of Mathematics and Statistics, University of Ottawa*

[4]*Princess Margaret Cancer Centre, University Health Network*

[5]*Department of Biostatistics, University Health Network*

[6]*Department of Statistical Sciences, University of Toronto*

ABSTRACT

Instrumental variables (IVs) are often continuous, arising in diverse fields including economics, epidemiology, and the social sciences. Existing approaches for continuous IVs typically impose strong parametric models or homogeneous treatment effects, while fully nonparametric methods may perform poorly in moderate to high dimensions. We propose a new framework for identifying and estimating the average treatment effect (ATE) with continuous IVs via the conditional weighted average derivative effect (CWADE). Using a conditional Riesz representation, our framework also unifies continuous and categorical IVs. In this framework, the ATE is typically overidentified, leading to a semiparametric observed data model with nonlinear constraints on the tangent space. Addressing these nonlinear constraints requires delicate tools: we develop semiparametric efficiency theory using a second-order parametric submodel, which, to the best of our knowledge, has not been standard practice in this literature. For estimation, we characterize a class of conditional reverse Riesz representers for the CWADE, yielding an easy-to-implement, triply robust estimator that is also locally efficient. We apply our methods to a novel dataset from the Princess Margaret Cancer Centre to examine the so- called obesity paradox in oncology, assessing the causal effect of excess body weight on two-year mortality among patients with non–small cell lung cancer.

Keywords: Mendelian randomization; Riesz representers; Semiparametric methods; Unmeasured confounding

Automated Analysis of Experiments using Hierarchical Garrote

Wei-Yang Yu and V. Roshan Joseph*

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332

ABSTRACT

In this work, we propose an automatic method for the analysis of experiments that incorporates hierarchical relationships between the experimental variables. We use a modified version of the nonnegative garrote method for variable selection which can incorporate hierarchical relationships. The nonnegative garrote method requires a good initial estimate of the regression parameters for it to work well. To obtain the initial estimate, we use generalized ridge regression with the ridge parameters estimated from a Gaussian process prior placed on the underlying input-output relationship. The proposed method, called HiGarrote, is fast, easy to use, and requires no manual tuning. Analysis of several real experiments are presented to demonstrate its benefits over the existing methods.

Keywords: Gaussian process; Generalized ridge regression; Nonnegative garrote; Variable selection

Uncertainty Quantification for Noisy Low-Tubal-Rank Tensor Completion

Jiuqian Shang, Jingyang Li, Yang Chen

Department of Statistics, University of Michigan

ABSTRACT

The low-tubal-rank tensor model has been used for multidimensional data to capture signals in the frequency domain. Algorithms have been developed to estimate low-rank third-order tensors from partial and corrupted entries. However, uncertainty quantification and statistical inference for these estimates remain largely unclear. We introduce a flexible framework for inference on general linear forms of a large tensor whenever an entry-wise consistent estimator is available. Under mild regularity conditions, we construct asymptotically normal estimators of these linear forms through double-sample debiasing and low-rank projection. These estimators allow us to construct confidence intervals and perform hypothesis testing. Simulation studies support our theoretical results. We apply the method to the total electron content (TEC) reconstruction problem, demonstrating that it delivers more robust reconstructions and informative entry-wise confidence intervals.

Keywords: Spectral method; Asymptotic normality; Total electron content (TEC)

Linearly Constrained Symmetric Rank-One Approximation for Pre-Image Recovery in Nonlinear Matrix Completion

Elvin Tseng[1], Laura Balzano[1,2], Gregory Ongie[3]

[1]*Department of Statistics, University of Michigan*

[2]*Department of Electrical Engineering & Computer Science, University of Michigan*

[3]*Department of Mathematical and Statistical Sciences, Marquette University*

ABSTRACT

We study the linearly constrained rank-one approximation of a symmetric matrix, motivated by pre-image recovery and structured matrix estimation. We show that the problem admits a simple geometric reduction to a one-dimensional equation whose largest root yields the global minimizer. Under mild assumptions, this solution is unique. We further give a verifiable condition that distinguishes true local minima from spurious stationary points. We also clarify connections to generalized eigenvalue problem and trust-region formulations, revealing strong but non-trivial structural links. For large-scale instances where the required eigendecomposition is computationally intensive, we develop a scalable projected-gradient method that avoids explicit factorization and remains effective in high dimensions.

Keywords: Constrained optimization; Generalized eigenvalue problem; Nonlinear matrix completion; Trust-region subproblem; Spurious local minima

Investigating the Pleiotropic Effects of Splice-Affecting Variants on Cancer and Metabolic Diseases

Jia-Ying Su[1], En-Yu Lai[1], Chien-Ling Lin[2], Yen-Tsung Huang[1,*]

[1]*Institute of Statistical Science, Academia Sinica*

[2]*Institute of Molecular Biology, Academia Sinica*

ABSTRACT

Genome-wide association studies (GWAS) have identified numerous loci for complex diseases, but a key challenge is to understand their biological functions. A specific class of potent functional variants, known as Splice-Affecting Variants (SAVs), alter the RNA splicing process, which can lead to the production of aberrant proteins and disrupt cellular function. There is substantial evidence that metabolic dysregulation, such as that seen in type 2 diabetes and obesity, is a significant risk factor for various cancers. Therefore, we aim to investigate whether SAVs confer risk for both metabolic diseases and cancer through a shared genetic mechanism, a phenomenon known as pleiotropy.

We leveraged genotype and electronic health record data from over 560,000 participants in the Taiwan Precision Medicine Initiative (TPMI). Our analysis focused on 468 high-confidence SAVs predicted by our previously developed tool, spliceAPP. We then employed a novel statistical framework designed to integrate evidence from two different traits. This method substantially boosts the power to identify shared genetic architecture, even when one of the signals is modest.

Our analyses indicate that several SAVs are significantly associated with the risk of metabolic diseases while also demonstrating a concurrent association with cancer risk (FDR < 0.01). Specifically, 118 SAVs were identified between prostate cancer and type 2 diabetes, while 71 SAVs were found between rectal cancer and type 2 diabetes. Among these, a core set of 38 SAVs was significant in both relationships. Notably, several of these shared variants are located in genes with well-established roles in both cancer and metabolic diseases, including *CDKN1A*, *TNFRSF1B*, *C2*, *SERINC2*, and *ST3GAL6*. These findings lend support to our pleiotropy hypothesis, offering new insights into how SAVs may drive cancer pathophysiology through shared biological pathways and highlighting the potential of integrative analytical strategies to unravel the genetic basis of complex diseases.

Keywords: Pleiotropy; Splice-Affecting Variants (SAVs); Cancer Genetics; Metabolic Disease; Taiwan Precision Medicine Initiative (TPMI).

Controlling the False Discovery Proportion in Observational Studies with Hidden Bias

Menggi Lin, Colin Fogarty

University of Michigan

ABSTRACT

We propose an approach to exploratory data analysis in matched observational studies. We consider the setting where a single intervention is thought to potentially impact multiple outcome variables, and the researcher would like to investigate which of these causal hypotheses come to bear while accounting not only for the possibility of false discoveries, but also the possibility that the study is plagued by unmeasured confounding. For any candidate set of rejected hypotheses, our method provides sensitivity sets for the false discovery proportion (FDP), the proportion of rejected hypotheses that are actually true. For a set \cR containing |\cR| outcomes, the method describes how much unmeasured confounding would need to exist for us to believe that the proportion of true hypotheses is 0/|\cR|, 1/|\cR|, ..., all the way to |\cR|/|\cR|. Moreover, the resulting confidence statements intervals are valid simultaneously over all possible choices for the rejected set, allowing the researcher to look in an ad hoc manner for promising subsets of outcomes that maintain a large estimated fraction of true discoveries even if a large degree of unmeasured confounding is present. The approach is particularly well suited to sensitivity analysis, as conclusions that some fraction of outcomes were affected by the treatment exhibit larger robustness to unmeasured confounding than the conclusion that any particular outcome was affected. In principle, the method requires solving a series of quadratically constrained integer programs. That said, we show not only that a solution can be obtained in reasonable run time, but also that one can avoid running the integer program altogether with high probability in large samples. We illustrate the practical utility of the method through simulation studies and a data example.

Keywords: Causal Inference; False Discovery Proportion; Matching; Sensitivity Analysis

LBBNN: An R package for Sparse Bayesian Neural Networks

Lars Skaaret-Lund, Eirik Høyheim, Aliaksandr Hubin

Bioinformatics and applied statistics, Norwegian University of Life Sciences

ABSTRACT

Artificial neural networks are highly popular and successful non-linear statistical models, used in a wide variety of applications. They do however have some drawbacks, mainly that they are heavily overparameterized, leading to overfitting. There are several ways of mitigating this, one is to incorporate parameter uncertainty through a Bayesian neural network (BNN). Model averaging also becomes straight-forward with BNNs, which typically improves performance. It is also of interest to reduce the amount of parameters of BNNs, to make them more interpretable, and reduce the memory footprint. This can be done in a principled way with latent binary Bayesian neural networks (LBBNN) [1], where a Bernoulli variable is placed in front of each weight in the network. While this incurs an extra parameter per weight compared to standard BNNs, the resulting sparse networks can make up for this. For example, both [1] and [2] demonstrate that networks with only 10 % of the original weights maintain similar accuracy and calibration metrics.

A further advance is to concatenate the input-layer to each hidden layer. Doing this allows for learning different functions based on the data. For example, in [3], we demonstrate that if we generate synthetic data that is linear, the network learns to remove all non-linear transformations and reduces to only a linear layer. This is in contrast to traditional statistical models where the assumptions on what function to model has to be made a priori. In [3], we also demonstrate that our method generates much sparser representations than standard LBBNNs, and also sparser than other related baselines, while maintaining high predictive power. Additionally, we introduce the concept of active paths, which is simply a path from an input node, either directly or via one or more hidden units to an output node. The motivation behind this is that pruning edges in a network may leave us with inactive nodes, meaning that no information going through these nodes will contribute to the output. This then can leave us with an even sparser representation than just considering the proportion of removed edges. Furthermore, the concept of active paths allows us to obtain both global and local explanations of predictions of the network. The goal of the R package LBBNN is to unify all the above mentioned approaches to make it usable for a wider audience. We allow the user to choose between standard LBBNNs, LBBNNs with normalizing flows (moving beyond mean-field Gaussian posteriors to allow for non-independent weights, as demonstrated in [2], and LBBNNs

with input-skip.

Keywords: Bayesian neural networks, variational inference, uncertainty, explainability

References:

[1] A. Hubin and G. Storvik. Sparse Bayesian neural networks: Bridging model and parameter uncertainty through scalable variational inference. Mathematics, 12(6):788, 2024.

[2] L. Skaaret-Lund, G. Storvik, and A. Hubin. Sparsifying Bayesian Neural Networks with Latent Binary Variables and Normalizing Flows. Transactions on Machine Learning Research, 10 2024. URL https://openreview.net/forum?id=d6kqUKzG3V.

[3] E. Høyheim, L. Skaaret-Lund, S. Sæbø, and A. Hubin. Explainable bayesian deep learning through input-skip latent binary bayesian neural networks. arXiv preprint arXiv:2503.10496, 2025

Back to Poster List

Cramér-Type Moderate Deviation and Berry–Esseen Bounds in the p-Spin Curie–Weiss Model

Tianyu Liu[1], Somabha Mukherjee[1], Bhaswar B. Bhattacharya[2]

[1]*Department of Statistics and Data Science, National University of Singapore, Singapore*

[2]*Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, USA*

ABSTRACT

Limit theorems for the magnetization in the p-spin Curie–Weiss model, for $p \geq 3$, have been derived recently. In this work, we strengthen these results by proving Cramér-type moderate deviation theorems and Berry–Esseen bounds for the magnetization (suitably centered and scaled). In particular, we show that the rate of convergence is on the order $O(N^{-1/2})$ when the magnetization has asymptotically Gaussian fluctuations, and $O(N^{-1/4})$ when the fluctuations are non-Gaussian. As an application, we derive a Berry–Esseen bound for the maximum pseudolikelihood estimate of the inverse temperature in the p-spin Curie–Weiss model with no external field, for all points in the parameter space where consistent estimation is possible.

Keywords: Central Limit Theorems; Stein's Method; Moderate Deviation; Spin Systems

Concentration-QTC Analysis to Support ICH E14 with a Real Case

Kong Xin[1], Xin Zhang[2]

Biostatistics, Sanofi

ABSTRACT

This presentation introduces ICH E14 and Concentration-QTC model. The PK and ECG data obtained from a Phase I study characterized the effects of the drug and had demonstrated there were no effects on QT/QTc intervals and other ECG parameters. The results were adequate to waive a formal TQT trial for regulatory submission.

Keywords: Concentration-QTC, Pharmacokinetics, ICH E14, Thorough QT/QTc, Early Phase

Participants

Masato Abe	Doshisha University	Invited Speaker
Elif Acar	University of Guelph	Invited Speaker
Joshua Agterberg	University of Illinois Urbana-Champaign	Invited Speaker
Raed Al Kontar	University of Michigan	Invited Speaker
Genevera Allen	Columbia University	Invited Speaker
Takumi Aoki	Pharmaceuticals and Medical Devices Agency	Invited Speaker
Jesus Arroyo	Texas A&M University	Invited Speaker
Koko Asakura	National Cerebral and Cardiovascular Center	Invited Speaker
Marco Avella Medina	Columbia University	Invited Speaker
Margaret Banker	Northwestern University Feinberg School of Medicine	Invited Speaker
Sumanta Basu	Cornell University	Invited Speaker
Claudio Giovanni Borroni	University of Milano - Bicocca	Invited Speaker
Frank Bretz	Novartis Pharma AG	Invited Speaker
Nicolas Brunel	ENSIIE	Session Organizer
Qiushi Bu	City University of Hong Kong	Invited Speaker
Biao Cai	City University of Hong Kong	Invited Speaker
Changxiao Cai	University of Michigan	Invited Speaker
Jianwen Cai	University of North Carolina at Chapel Hill	Invited Speaker
Mingxuan Cai	City University of Hong Kong	Invited Speaker
Zhanrui Cai	University of Hong Kong	Invited Speaker
Hongyuan Cao	Florida State University	Invited Speaker
Manuela Cazzaro	University of Milano - Bicocca	Invited Speaker
Kwun Chuen Gary Chan	University of Washington	Invited Speaker
Chih-Yu Chang	Imperial College London	General Participant
Hsin-wen Chang	Institute of Statistical Science, Academia Sinica	Invited Speaker
I-Shou Chang	National Health Research Institutes, Taiwan	General Participant
Joyce Chang	University of Pittsburgh	Invited Speaker
Kyle Chang	Micron Technology	General Participant
Lun-Ching Chang	Florida Atlantic University	General Participant
Ming-Chung Chang	Institute of Statistical Science, Academia Sinica	Invited Speaker
Wei Shan Chang	Institute of Statistical Science, Academia Sinica	General Participant
Wei-hau Chang	Institute of Statistical Science, Academia Sinica	Invited Speaker
Yuan-chin Chang	Institute of Statistical Science, Academia Sinica	General Participant

Yu-Wei Chang	National Chengchi University	General Participant
Carolin Chao	Parexel International Co., Ltd	General Participant
Andrew Chen	Medical University of South Carolina	General Participant
Bo-yu Chen	Purdue University	General Participant
Cathy Woan Shu Chen	Feng Chia University	Invited Speaker
Chen Fang Chen	Center for Drug Evaluation, Taiwan	General Participant
Chen-Hsin Chen	Institute of Statistical Science, Academia Sinica	General Participant
Chin-Chun Chen	National Cheng Kung University	General Participant
Chun-houh Chen	Institute of Statistical Science, Academia Sinica	Session Chair
Chun-Shu Chen	National Central University	Session Organizer
Chyong-Mei Chen	National Yang Ming Chiao Tung University	Invited Speaker
George Chen	Carnegie Mellon University	Invited Speaker
Hao Chen	University of California, Davis	Invited Speaker
Hsuanyu Chen	Institute of Statistical Science, Academia Sinica	Invited Speaker
I-Hsiao Chen	National Tsing Hua University	General Participant
Jiahua Chen	The University of British Columbia	Invited Speaker
Jie Chen	Taimei Intelligence Biopharma R&D	Invited Speaker
Jyun-Yu Chen	Institute of Statistical Science, Academia Sinica	General Participant
Kani Chen	The Hong Kong University of Science and Technology	Invited Speaker
Li-Pang Chen	National Chengchi University	Invited Speaker
Manhua Chen	Tamkang University	General Participant
May-Ru Chen	National Sun Yat-sen University	Invited Speaker
Nan Chen	National University of Singapore	Invited Speaker
Ping-Yang Chen	National Taipei University	General Participant
Qingxia Chen	Vanderbilt University Medical Center	Invited Speaker
Ray-Bing Chen	National Tsing Hua University	Invited Speaker
Rong Chen	Rutgers University	General Participant
Shizhe Chen	University of California, Davis	Invited Speaker
Shu-Chuan Chen	Idaho State University	General Participant
Shu-Chun Chen	Institute of Statistical Science, Academia Sinica	General Participant
Ting-Li Chen	Institute of Statistical Science, Academia Sinica	General Participant
Tsung-Wen Chen	Institute of Statistical Science, Academia Sinica	General Participant
Wan Ping Chen	Fu Jen Catholic University	General Participant
Wei Chen	University of Pittsburgh	Invited Speaker
Wilson Ye Chen	University of Sydney	Invited Speaker
Yan-Bin Chen	National Taiwan University	Session Chair
Yan-Lin Chen	National Yang Ming Chiao Tung University	General Participant

Yaqing Chen	Rutgers University	Invited Speaker
Yi Chen	Hong Kong University of Science and Technology	Invited Speaker
Yuan Chen	Memorial Sloan Kettering Cancer Center	Invited Speaker
Yu-Wei Chen	Purdue University	General Participant
Zih-Bing Chen	Iowa State University	General Participant
Chi-Lun Cheng	Institute of Statistical Science, Academia Sinica	General Participant
Hao-Chung Cheng	National Taiwan University	Invited Speaker
Yu-Jen Cheng	National Tsing Hua University	Invited Speaker
Chien-Ming Chi	Institute of Statistical Science, Academia Sinica	General Participant
Yuejie Chi	Yale University	Invited Speaker
Kuo-Szu Chiang	National Chung Hsing University	General Participant
Min-Hsiu Chiang	Institute of Statistical Science, Academia Sinica	General Participant
Wei-Chu Chiang	Institute of Statistical Science, Academia Sinica	General Participant
Yi Ting Chiang	Institute of Statistical Science, Academia Sinica	General Participant
Yen-Shiu Chin	Institute of Statistical Science, Academia Sinica	Invited Speaker
Jeng-Min Chiou	Institute of Statistical Science, Academia Sinica	Session Chair
Chun-Chieh Chiu	Institute of Statistical Science, Academia Sinica	General Participant
Chen-Hua Cho	National Tsing Hua University	General Participant
Shu-Hsien Cho	University of Texas MD Anderson Cancer Center	General Participant
Young Hyun Cho	Purdue University	Invited Speaker
Sung-Hyuk Choi	Korea University Guro Hospital	General Participant
Boonyarit Choopradit	Thammasat University	General Participant
Boris Choy	The University of Sydney	Session Organizer
Amanda Chu	The Education University of Hong Kong	General Participant
Lynna Chu	Iowa State University	Invited Speaker
Chia Jen Chuang	National Cheng Kung University	General Participant
Chien-Wei Chuang	Fu Jen Catholic University	Invited Speaker
Ching-Yen Chuang	National Tsing Hua University	General Participant
Szu-Chi Chung	National Sun Yat-sen University	Invited Speaker
Marc Coram	Google Zurich	Invited Speaker
Noel Cressie	University of Wollongong	Invited Speaker
Erjia Cui	University of Minnesota	General Participant
Xinping Cui	University of California, Riverside	Session Organizer
Ben Dai	The Chinese University of Hong Kong	Invited Speaker
Chi-Shian Dai	National Cheng Kung University	Invited Speaker
Ran Dai	University of Nebraska Medical Center	Invited Speaker
Xiaowu Dai	University of California, Los Angeles	Invited Speaker

Milan Kumar Das	Institute of Statistical Science, Academia Sinica	General Participant
Carlos de la Calle-Arroyo	Universidad de Oviedo	Invited Speaker
Vaibhav Dhameliya	Sardar Vallabhbhai National Institute of Technology, Surat, India	General Participant
Jimin Ding	Washington University in St. Louis	Invited Speaker
Xiucai Ding	University of California, Davis	Invited Speaker
Ying Ding	University of Pittsburgh	Session Organizer
Mei Dong	University of Toronto	General Participant
Xiaoling Dou	International Christian University	Invited Speaker
Congyuan Duan	Hong Kong University of Science and Technology	General Participant
Paromita Dubey	University of Southern California	Invited Speaker
Saykat Dutta	Institute of Statistical Science, Academia Sinica	General Participant
Abdel El-shaarawi	Cairo university	Invited Speaker
Takeshi Emura	Hiroshima University	Invited Speaker
Sylvia Esterby	University of British Columbia	Session Chair
Jianqing Fan	Princeton University	Invited Speaker
Jiarong Fan	University Paris-Saclay	Invited Speaker
Yingying Fan	University of Southern California	Invited Speaker
Zhaozhi Fan	Memorial University of Newfoundland	General Participant
Wei-Quan Fang	Center for Drug Evaluation, Taiwan	General Participant
Shai Feldman	Technion	Invited Speaker
Hao Feng	The University of Texas Health Science Center at Houston	Invited Speaker
Zeny Feng	University of Guelph	Session Organizer
Hui Mean Foo	Institute of Statistical Science, Academia Sinica	General Participant
Haoda Fu	Amgen Inc.	General Participant
Erin Gabriel	University of Copenhagen	Invited Speaker
Etienne Gauthier	University Inria	Invited Speaker
Yu Gu	University of Hong Kong	Invited Speaker
Meihui Guo	National Sun Yat-sen University	Invited Speaker
Xinzhou Guo	The Hong Kong University of Science and Technology	Invited Speaker
Il Do Ha	Pukyong National University	Invited Speaker
Wooseok Ha	Korea Advanced Institute of Science and Technology	Invited Speaker
Toshimitsu Hamasaki	The George Washington University	Invited Speaker
Fang Han	University of Washington	Invited Speaker
Wei Hao	University of Michigan	Invited Speaker
Wenqing He	University of Western Ontario	Invited Speaker
Xin He	University of Maryland, College Park	Invited Speaker
Masaki Hino	The Graduate University for Advanced Studies, SOKENDAI	Invited Speaker

Ho Yi Alexis Ho	Hong Kong University of Science and Technology	General Participant
Andrew Holbrook	University of California, Los Angeles	Session Organizer
Miki Horiguchi	Dana-Farber Cancer Institute	Invited Speaker
Chin-Fu Hsiao	National Health Research Institutes, Taiwan	Session Chair
Chao Hsiung	National Health Research Institutes, Taiwan	General Participant
Chih-Yuan Hsu	Vanderbilt University Medical Center	General Participant
Nan-Jung Hsu	National Tsing-Hua University	Invited Speaker
Wei-Chin Hsu	Institute of Statistical Science, Academia Sinica	General Participant
Guanyu Hu	Michigan State University	Invited Speaker
Jie Hu	The Ohio State University	Invited Speaker
Robert Hsuan-Fu Hua	Florida State University	General Participant
Sheng-Zhan Hua	University of California, Los Angeles	General Participant
Chao-Hui Huang	Institute of Statistical Science, Academia Sinica	General Participant
Chia-Hui Huang	National Chengchi University	Invited Speaker
Chiung-Yu Huang	University of California, San Francisco	Invited Speaker
Guan-Hua Huang	National Yang Ming Chiao Tung University	General Participant
Hsin-Cheng Huang	Institute of Statistical Science, Academia Sinica	General Participant
Hsueh-Han Huang	Institute of Statistical Science, Academia Sinica	Invited Speaker
Jing-Wen Huang	Institute of Statistical Science, Academia Sinica	Invited Speaker
Li-Shan Huang	National Tsing Hua University	Invited Speaker
Ming-Yueh Huang	Institute of Statistical Science, Academia Sinica	General Participant
Shih-Feng Huang	National Central University	General Participant
Shih-Hao Huang	National Central University	Session Chair
Shih-Ting Huang	University of Louisville	General Participant
Shuo-Chieh Huang	Rutgers University	Invited Speaker
Su-Yun Huang	Institute of Statistical Science, Academia Sinica	Session Organizer
Wei-Heng Huang	National Taipei University	General Participant
Whitney Huang	Clemson University	Invited Speaker
Yiting Huang	Institute of Statistical Science, Academia Sinica	General Participant
Hung Hung	National Taiwan University	Invited Speaker
Ying-Chao Hung	National Taiwan University	General Participant
Jing-Shiang Hwang	Institute of Statistical Science, Academia Sinica	General Participant
Wei-Ting Hwang	University of Pennsylvania	General Participant
Wen-Han Hwang	National Tsing Hua University	General Participant
Ching-Kang Ing	National Tsing Hua University	Session Organizer
Nilah Ioannidis	University of California, Berkeley	Invited Speaker
Thomas Jensen	Utah State University	Invited Speaker

Yuan Ji	University of Chicago	Invited Speaker
Binyan Jiang	The Hong Kong Polytechnic University	Invited Speaker
Ci-Ren Jiang	National Taiwan University	Session Chair
Hui Jiang	University of Michigan	General Participant
Jiming Jiang	University of California, Davis	Invited Speaker
Yuchao Jiang	Texas A&M University	Invited Speaker
Shi Jin	Shanghai Jiao Tong University	Invited Speaker
Zhezhen Jin	Columbia University	Invited Speaker
Jinyeon Jo	Institute of Statistical Science, Academia Sinica	General Participant
Zhi-Yu Jou	Institute of Statistical Science, Academia Sinica	General Participant
Yuan-Jung Juang	University of California, Davis	Invited Speaker
Jian Kang	University of Michigan	Invited Speaker
Sangwook Kang	Yonsei University	Invited Speaker
Wei Chun Kang	National Tsing Hua University	General Participant
Ming-Hung (Jason) Kao	Arizona State University	Invited Speaker
Sayar Karmakar	University of Florida	Invited Speaker
Shogo Kato	Institute of Statistical Mathematics	Invited Speaker
Hyunjoong Kim	Yonsei University	General Participant
Inyoung Kim	Virginia Tech	Invited Speaker
Jihoon Kim	Pukyong University	General Participant
Kwangho Kim	Korea University	Invited Speaker
Mi-Ok Kim	University of California San Francisco	Invited Speaker
Myungjin Kim	Kyungpook National University	Invited Speaker
Seoung Bum Kim	Korea University	General Participant
Takaaki Koike	Hitotsubashi University	Invited Speaker
Maiying Kong	University of Louisville	Session Organizer
Samuel Kou	Harvard University	Invited Speaker
Meifen Kung	Center for Drug Evaluation, Taiwan	General Participant
Chiishyang Kuo	Institute of Statistical Science, Academia Sinica	General Participant
En-Yu Lai	Institute of Statistical Science, Academia Sinica	Contributed Poster
J. Jack Lee	University of Texas MD Anderson Cancer Center	General Participant
Jeseok Lee	Kyungpook national University	General Participant
Kuo-Jung Lee	National Cheng Kung University	General Participant
Mei-Ling Ting Lee	University of Maryland	Session Organizer
Sanghyeok Lee	Small Enterprise and Market Service	General Participant
Yi-Ju (Jean) Lee	Institute of Statistical Science, Academia Sinica	General Participant
Cheng-Shiun Leu	Columbia University	Invited Speaker

Degui Li	University of Macau	Invited Speaker
Fan Li	Duke University	Invited Speaker
Fan Li	Yale University	Invited Speaker
Gang Li	University of California, Los Angeles	Invited Speaker
Huilin Li	New York University	Invited Speaker
Jialiang Li	National University of Singapore	Invited Speaker
Jie Li	Renmin University of China	General Participant
Jingyang Li	University of Michigan	Invited Speaker
Qian Li	StatsVita, LLC	Invited Speaker
Qiwei Li	The University of Texas at Dallas	Invited Speaker
Ting Li	Hong Kong Polytechnic University	Session Chair
Xiaodong Li	University of California, Davis	Invited Speaker
Xinran Li	University of Chicago	Invited Speaker
Yi Li	University of Michigan	Invited Speaker
Yu-Cheng Li	Institute of Statistical Science, Academia Sinica	General Participant
Yu-Hsi Li	Institute of Statistical Science, Academia Sinica	General Participant
Yun Li	University of North Carolina	Invited Speaker
Zhaoyuan Li	The Chinese University of Hong Kong, Shenzhen	General Participant
C. Jason Liang	National Institute of Allergy and Infectious Diseases	Session Organizer
Feng Liang	University of Illinois Urbana-Champaign	Invited Speaker
Yu-Jen Liang	Institute of Statistical Science, Academia Sinica	General Participant
Cai-Sian Liao	Institute of Statistical Science, Academia Sinica	General Participant
Xiyue Liao	San Diego State University	General Participant
Yu-Hsiang Lien	Institute of Statistical Science, Academia Sinica	General Participant
Chae Young Lim	Seoul National University	Invited Speaker
Chang-Yun Lin	National Chung Hsing University	Invited Speaker
Chia-Wei Lin	National Tsing Hua University	General Participant
Chien-Tai Lin	Tamkang University	General Participant
Chin-Yi Lin	Foxconn	Invited Speaker
Danyu Lin	University of North Carolina	Invited Speaker
Dennis Lin	Purdue University	Invited Speaker
Feng-Chang Lin	University of North Carolina at Chapel Hill	Session Organizer
Gwo Dong Lin	Institute of Statistical Science, Academia Sinica	General Participant
Hsin Hung Lin	Institute of Statistical Science, Academia Sinica	General Participant
Liang-Ching Lin	National Cheng Kung University	General Participant
Pei-Rou Lin	Institute of Statistical Science, Academia Sinica	General Participant
Shih-yu Lin	Institute of Statistical Science, Academia Sinica	General Participant

Shu-Chin Lin	National Taiwan University	Invited Speaker
Szu-Han Lin	Institute of Statistical Science, Academia Sinica	General Participant
Tzu-Chuan Lin	Center for Drug Evaluation, Taiwan	Invited Speaker
Xihong Lin	Harvard University	Invited Speaker
Yuanyuan Lin	The Chinese University of Hong Kong	Invited Speaker
Zhenhua Lin	National University of Singapore	Invited Speaker
Zhixiang Lin	The Chinese University of Hong Kong	Invited Speaker
Michelle Liou	Institute of Statistical Science, Academia Sinica	General Participant
Catherine Liu	The Hong Kong Polytechnic University	Invited Speaker
ChunFan Liu	Center for Drug Evaluation, Taiwan	General Participant
Fang Liu	University of Notre Dame	Invited Speaker
Huai-Chien Liu	National Taiwan Normal University	General Participant
Jingbo Liu	University of Illinois	Invited Speaker
Kin Yat Liu	The Chinese University of Hong Kong	Invited Speaker
Lin Liu	Shanghai Jiao Tong University	Invited Speaker
Qingfeng Liu	Hosei University	Invited Speaker
Thomas Liu	Amgen Inc.	Invited Speaker
Wei-chung Liu	Institute of Statistical Science, Academia Sinica	Invited Speaker
Yan Liu	Waseda University	Invited Speaker
Zhonghua Liu	Columbia University	Invited Speaker
Yang Lo	Tunghai University	General Participant
Yun-Shuo Lo	Institute of Statistical Science, Academia Sinica	General Participant
Mong-Na Lo Huang	National Sun Yat-sen University	Invited Speaker
Qi Long	University of Pennsylvania	Invited Speaker
Jesús López-Fidalgo	Univertity of Navarra	Invited Speaker
Henry Horng-Shing Lu	Kaohsiung Medical University and	Invited Speaker
Hsien-Ting Lu	National Taiwan University	General Participant
Hsinyang Lu	National Taiwan University	General Participant
Qiongshi Lu	University of Wisconsin-Madison	Invited Speaker
Xinyi Lu	Utah State University	Invited Speaker
Ying Lu	Stanford University	Invited Speaker
Yuanhang Luo	Hong Kong Polytechnic University	Invited Speaker
Jinchi Lv	University of Southern California	Invited Speaker
Kwan-Liu Ma	University of California at Davis	Invited Speaker
Pulong Ma	Iowa State University	Invited Speaker
Shuangge Ma	Yale University	Invited Speaker
Tianzhou Ma	University of Maryland	Invited Speaker

Anil Maurya	Sardar Vallabhbhai National Institute of Technology, Surat, India	General Participant
Hao Mei	Renmin University of China	Invited Speaker
Xiaoli Meng	Harvard University	Invited Speaker
George Michailidis	University of California, Los Angeles	Session Organizer
Caleb Miles	Columbia University	Invited Speaker
Kevin Moon	Utah State University	Invited Speaker
Toshinari Morimoto	National Taiwan University	General Participant
Satoshi Morita	Kyoto University Graduate School of Medicine	Invited Speaker
Jonas Mueller	Cleanlab	Invited Speaker
Amitava Mukherjee	XLRI - Xavier School of Management	Invited Speaker
Hans-Georg Müller	University of California, Davis	Invited Speaker
Thuan Nguyen	Oregon Health and Science University	Invited Speaker
Jing Ning	The University of Texas M.D. Anderson Cancer Center	Invited Speaker
Ning Ning	Texas A&M University	Invited Speaker
Todd Ogden	Columbia University	Invited Speaker
Shuhei Ota	Kanagawa University	Session Organizer
Fang-Shu Ou	Mayo Clinic	General Participant
Wei Pan	University of Minnesota	Invited Speaker
Byeong Uk Park	Seoul National University	Invited Speaker
Frederick Kin Hing Phoa	Institute of Statistical Science, Academia Sinica	Invited Speaker
Tseng Po Shan	Tsinghua University	General Participant
Aman Prakash	Sardar Vallabhbhai National Institute of Technology, Surat, India	General Participant
Soumik Purkayastha	University of Pittsburgh	Invited Speaker
Peihua Qiu	University of Florida	Session Chair
Hsiau Ren	National Changhua University of Education	Invited Speaker
Kun Ren	City University of Hong Kong	Invited Speaker
Zhao Ren	University of Pittsburgh	Invited Speaker
Juan M. Rodriguez-Diaz	University of Salamanca	Invited Speaker
Michael Sachs	University of Copenhagen	Invited Speaker
Srijan Sengupta	North Carolina State University	Invited Speaker
Jiuqian Shang	University of Michigan	General Participant
Jun Shao	University of Wisconsin	Invited Speaker
Dennis Shen	University of Southern California	Invited Speaker
Guohao Shen	The Hong Kong Polytechnic University	Invited Speaker
Hua Shen	University of Calgary	Session Chair
Chenlu Shi	New Jersey Institute of Technology	Invited Speaker
Shota Shibasaki	Doshisha University	Invited Speaker

Shwu-Rong Shieh	Institute of Statistical Science, Academia Sinica	General Participant
Yun-Jer Shieh	Institute of Statistical Science, Academia Sinica	General Participant
Jia-Han Shih	National Sun Yat-sen University	Invited Speaker
Yei Eun Shin	Seoul National University	Invited Speaker
Hai Shu	New York University	Invited Speaker
Jeng Shuen-Lin	National Cheng Kung University	Invited Speaker
Huei-Lun Siao	Institute of Statistical Science, Academia Sinica	General Participant
CY (Chor-yiu) Sin	National Tsing Hua University	General Participant
Tony Sit	The Chinese University of Hong Kong	Invited Speaker
Lars Skaaret-Lund	Norwegian University of Life Sciences	General Participant
Mike So	The Hong Kong University of Science and Technology	General Participant
Peter Song	University of Michigan	Invited Speaker
Xinyuan Song	The Chinese University of Hong Kong	Invited Speaker
John Stevens	Utah State University	Invited Speaker
John Stufken	George Mason University	Invited Speaker
Jia-Ying Su	Institute of Statistical Science, Academia Sinica	Contributed Poster
Pei-Fang Su	National Cheng Kung University	Session Chair
Yoshiki Sugimoto	Doshisha University	Invited Speaker
Cheng-Yu Sun	National Tsing Hua University	Invited Speaker
Jiayang Sun	George Mason University	Invited Speaker
Li-Hsien Sun	National Central University	General Participant
Saifei Sun	City University of Hong Kong	General Participant
Wei-Hsiang Sun	University of Michigan, Ann Arbor	General Participant
An-Shun Tai	National Tsing Hua University	General Participant
Yi-Cheng Tai	National Chengchi University	General Participant
Nanami Taketomi	Nagasaki University	Invited Speaker
Ming Tan	Georgetown University	General Participant
Hua Tang	Stanford University	Invited Speaker
Rong Tang	Hong Kong University of Science and Technology	Invited Speaker
Szu-Yu Tang	Pfizer Inc.	Invited Speaker
Chin-Sheng Teng	University of California, Riverside	General Participant
Huei-Wen Teng	National Yang Ming Chiao Tung University	General Participant
Lu Tian	Stanford University	Invited Speaker
Naitee Ting	StatsVita, LLC	Invited Speaker
Khong-Loon Tiong	Institute of Statistical Science, Academia Sinica	General Participant
Guangyu Tong	Yale University	Invited Speaker
Xin Tong	National University of Singapore	Invited Speaker

Arthur Tsai	Institute of Statistical Science, Academia Sinica	Invited Speaker
Guei-Feng (Cindy) Tsai	Center for Drug Evaluation, Taiwan	Session Organizer
Henghsiu Tsai	Institute of Statistical Science, Academia Sinica	Session Organizer
Katherine Tsai	Apple	Invited Speaker
Ruey Tsay	National Tsing Hua University	General Participant
Elvin Tseng	University of Michigan	General Participant
George Tseng	University of Pittsburgh	Invited Speaker
Panagiotis Tsiamyrtzis	Politecnico di Milano	Invited Speaker
I-Ping Tu	Institute of Statistical Science, Academia Sinica	Session Organizer
Hung-Ping Tung	National Yang Ming Chiao Tung University	General Participant
Hajime Uno	Harvard Medical School/Dana-Farber Cancer Institute	Invited Speaker
Chao Wang	The University of Sydney	Invited Speaker
Chen Wang	The University of Hong Kong	General Participant
Chiatse Wang	Institute of Statistical Science, Academia Sinica	General Participant
Chien-Chung Wang	Colorado State University	General Participant
Ci-Yu Wang	University of Michigan	General Participant
HaiYing Wang	University of Connecticut	Invited Speaker
Huixia Judy Wang	Rice University	Session Organizer
Jane-Ling Wang	University of California, Davis	Session Organizer
Jiebiao Wang	University of Pittsburgh	Invited Speaker
Jingshu Wang	The University of Chicago	Invited Speaker
Junhui Wang	The Chinese University of Hong Kong	Invited Speaker
Lei Wang	The Lotus Group	Invited Speaker
Lin Wang	Purdue University	Invited Speaker
Linbo Wang	University of Toronto	Invited Speaker
Mey Wang	Institute of Statistical Science, Academia Sinica	General Participant
Naisyin Wang	University of Michigan	General Participant
ShaoHsuan Wang	National Central University	Session Chair
Shulei Wang	University of Illinois Urbana-Champaign	General Participant
Tzu-Chi Wang	Institute of Statistical Science, Academia Sinica	General Participant
Wanjie Wang	National University of Singapore	Invited Speaker
Weijing Wang	National Yang Ming Chiao Tung University	General Participant
Weining Wang	University of Bristol	Invited Speaker
Wenyang Wang	Dalian Maritime University	Invited Speaker
Xia Wang	University of Cincinnati	Invited Speaker
Xiaofeng Wang	Cleveland Clinic	Invited Speaker
Yi-Ting Wang	National Taiwan University	Invited Speaker

Yinghui Wei	University of Plymouth	General Participant
Yingying Wei	The Chinese University of Hong Kong	Invited Speaker
Yun Ting Wei	Fu Jen Catholic University	General Participant
Nuttanan Wichitaksorn	Auckland University of Technology	Invited Speaker
Ka-Chun Wong	City University of Hong Kong	Invited Speaker
Kin Yau Wong	The Hong Kong Polytechnic University	Invited Speaker
Weng Kee Wong	University of California, Los Angeles	Invited Speaker
William WL Wong	University of Waterloo	Invited Speaker
Changbao Wu	University of Waterloo	Invited Speaker
Di Wu	University of North Carolina at Chapel Hill	General Participant
Han-Ming Wu	National Chengchi University	Invited Speaker
Hulin Wu	University of Texas Health Science Center at Houston	Invited Speaker
Jer-Yuarn Wu	Academia Sinica	Invited Speaker
Jia-Syuan Wu	Institute of Statistical Science, Academia Sinica	General Participant
Jingwei Wu	Temple University	General Participant
Kai-Yuan Wu	Institute of Statistical Science, Academia Sinica	Contributed Poster
Samuel Wu	University of South Florida	Invited Speaker
Shan-Chi Wu	Institute of Statistical Science, Academia Sinica	General Participant
Wayne Yi-Hung Wu	UTHealth/ MD Anderson Cancer Center	General Participant
Yichao Wu	University of Illinois Chicago	Invited Speaker
Yihong Wu	Yale University	Invited Speaker
Ying Nian Wu	University of California, Los Angeles	Invited Speaker
Zhijin Wu	Brown University	Invited Speaker
Amy Xia	Amgen Inc.	Invited Speaker
Dong Xia	Hong Kong University of Science and Technology	Invited Speaker
Liming Xiang	Nanyang Technological University	Invited Speaker
Hongquan Xu	University of California, Los Angeles	Session Organizer
Yanxun Xu	Johns Hopkins University	Invited Speaker
Haoran Xue	City University of Hong Kong	General Participant
Bao-Zhu Yang	Yale University	Invited Speaker
Hsin-Chou Yang	Institute of Statistical Science, Academia Sinica	Session Organizer
Jingyuan Yang	AbbVie	Session Chair
Jingzhen Yang	Nationwide Children's Hospital	Invited Speaker
Junho Yang	Institute of Statistical Science, Academia Sinica	Invited Speaker
Kang-Chung Yang	Institute of Statistical Science, Academia Sinica	General Participant
Min Yang	University of Illinois Chicago	Invited Speaker
Shihao Yang	Georgia Institute of Technology	Invited Speaker

Qiwei Yao	London School of Economics	Invited Speaker
Kazuyoshi Yata	University of Tsukuba	Invited Speaker
Jingjing Ye	BeOne Medicines	Invited Speaker
Chen-Hsiang Yeang	Institute of Statistical Science, Academia Sinica	Invited Speaker
Arthur Yeh	Bowling Green State University	Invited Speaker
ChiChun Yeh	Institute of Statistical Science, Academia Sinica	Contributed Poster
Grace Yi	University of Western Ontario	Invited Speaker
Guosheng Yin	University of Hong Kong	General Participant
Naoki Yoshimaru	Doshisha University	Invited Speaker
Guo Yu	University of California Santa Barbara	Invited Speaker
Jih-Chang Yu	National Taipei University	General Participant
Menggang Yu	University of Michigan	Invited Speaker
Ruoqi Yu	University of Illinois Urbana-Champaign	Invited Speaker
Wei-Yang Yu	Georgia Institute of Technology	General Participant
Ying Yuan	University of Texas MD Anderson Cancer Center	Invited Speaker
Chongzhi Zang	University of Virginia	Invited Speaker
Yong Zang	Indiana University	Session Organizer
Bin Zhang	Cincinnati Children's Hospital Medical Center	General Participant
Elio Zhang	Seattle University	Invited Speaker
Emma Jingfei Zhang	Emory University	Invited Speaker
Hao Zhang	Michigan State University	Session Organizer
Jin-Ting Zhang	National University of Singapore	General Participant
Min Zhang	University of California, Irvine	Invited Speaker
Nanhua Zhang	University of Cincinnati / Cincinnati Children's Hospital Medical Center	Invited Speaker
Tingting Zhang	University of Pittsburgh	Session Organizer
Yichen Zhang	Purdue University	Invited Speaker
Ying Zhang	University of Nebraska Medical Center	Invited Speaker
Zhihan Zhang	East China Normal University	Invited Speaker
Hongyu Zhao	Yale University	Invited Speaker
Yi Zhao	Indiana University School of Medicine	Invited Speaker
Yichuan Zhao	Georgia State University	Invited Speaker
Yunpeng Zhao	Colorado State University	Invited Speaker
Zhongming Zhao	University of Texas Health Science Center at Houston	Invited Speaker
Cheng Zheng	University of Nebraska Medical Center	Invited Speaker
Qi Zheng	University of Louisville	Invited Speaker
Tian Zheng	Columbia University	Session Chair

Wei Zheng	University of Tennessee	Invited Speaker
Haibo Zhou	University of North Carolina at Chapel Hill	Invited Speaker
Hua Zhou	University of California, Los Angeles	Invited Speaker
Xiang Zhou	Yale University	Invited Speaker
Changbo Zhu	University of Notre Dame	Invited Speaker
Ji Zhu	University of Michigan	Invited Speaker
Tianming Zhu	Nanyang Technological University	Invited Speaker
Xiaofeng Zhu	Case Western Reserve University	Invited Speaker
Yuhua Zhu	University of California, Los Angeles	Invited Speaker
Zhengyuan Zhu	Iowa State University	Invited Speaker
Li-Sheng Zhuang	Institute of Statistical Science, Academia Sinica	Contributed Poster
Baiming Zou	University of North Carolina at Chapel Hill	Invited Speaker
Hui Zou	University of Minnesota	Invited Speaker

JOINT 2025

https://www3.stat.sinica.edu.tw/joint2025/

Institute of Statistical Science, Academia Sinica

International Chinese Statistical Association

The Chinese Institute of Probability and Statistics