2024 International Conference for Statistics and Data Science

2024 統計與資料科學國際研討會

July 9-10, 2024

Agenda & Book of Abstracts

EXPLORE THE DATA

ORGANIZERS

ICSA 泛華統計協會

Table of Contents

Organizers & Committee

2024 International Conference for Statistics and Data Science, ICSDS

2024 統計與資料科學國際研討會

Date：July 9-10, 2024

Venue：College of Commerce, National Chengchi University

Organizers：

Department of Statistics, National Chengchi University

Institute of Statistics, National Yang Ming Chiao Tung University

Taiwan Chapter, International Chinese Statistical Association

Co-organizer / Sponsors：

Institute of Statistical Science, Academia Sinica

(中央研究院統計科學研究所)

NSTC Department of International Cooperation and Science Education

(國科會國合處)

SPEC Division of Mathematics

(科學推展中心數學組)

NCCU & College of Commerce

(國立政治大學&商學院)

National Yang Ming Chiao Tung University

(國立陽明交通大學)

Department of International and Cross-strait Education

(教育部國際及兩岸教育司)

Chinese Statistical Association (Taiwan)

(中國統計學社)

The Chinese Institute of Probability and Statistics

(中華機率統計學會)

NCCU Risk and Insurance Research Center

(政大風險與保險研究中心)

Department of Mathematical Sciences, NCCU

(國立政治大學應用數學系)

TransGlobe Life Insurance Inc.

(全球人壽保險股份有限公司)

Advisory Committee：

Samuel Kou, Harvard University
Mei-Ling Ting Lee, University of Maryland
Xiaodong Shen, University of Minnesota
Ruey S. Tsay, University of Chicago
Jane-Ling Wang, University of California
Martin T. Wells, Cornell University
Weng Kee Wong, University of California, Los Angeles
Hongyu Zhao, Yale University
Wendy Lou, University of Toronto
Liqun Wang, University of Manitoba
Bai-Yau Yeh, Bowling Green State University
Gang Li, UCLA Fielding School of Public Health
Zhezhen Jin, Columbia University
Xun Chen, Global Head of Biostatistics & Programming, Sanofi

Scientific Organizing Committee:

Henry Horng-Shing Lu (Chair), National Yang-Ming Chiao Tung University
Su-Fen Yang, National Chengchi University
Tsung-Chi Cheng, National Chengchi University
Guan-Hua Huang, National Yang-Ming Chiao Tung University
Yuan-chin Ivan Chang, Academia Sinica
I-Ping Tu, Academia Sinica
Jeng-Min Chiou, National Taiwan University
Li-Yu Daisy Liu, National Taiwan University
Hung Hung, National Taiwan University
Nan-Cheng Su, National Taipei University
Lee-Shen Chen, Ming Chuan University
Hung-Yi Lu, Fu Jen Catholic University
Shih-Feng Huang, National Central University
Chun-Shu Chen, National Central University
Wen-Han Hwang, National Tsing Hua University
Li-Shan Huang, National Tsing Hua University
Chao A. Hsiung, National Health Research Institutes
Tsung-I Lin, National Chung Hsing University
Yu-Mei Chang, Tunghai University
Cathy W.S. Chen, Feng Chia University
Juan-Ming Yuan, Providence University
Pei-Fang Su, National Cheng Kung University
Hsiang-Ling Hsu, National University of Kaohsiung

Mei-Hui Guo, National Sun Yat-Sen University
Wei-Ying Wu, National Dong Hwa University

Local Organizing Committee：

Su-Fen Yang (楊素芬) (Conference Chair and CEO)
Tsung-Chi Cheng (鄭宗記) (Vice CEO)
Chih-Hao Chang (張志浩)
Yu-Wei Chang (張育瑋)
Li-Pang Chen (陳立榜)
Vivian Yi-Ju Chen (陳怡如)
Yu-Ting Cheng (鄭宇庭)
Elizabeth Pei Ting Chou (周珮婷)
Huey-Miin Hsueh (薛慧敏)
Tzee-Ming Huang (黃子銘)
Chia-Hui Huang (黃佳慧)
Huimei Liu (劉惠美)
Ching-Syang Jack Yue (余清祥)
Chiu-Hsing Weng (翁久幸)
Han-Ming Wu (吳漢銘)

General Information

Traffic Guidance and Campus Maps

From Taiwan Taoyuan International Airport to NCCU

By Taxi

There is a taxi pickup area by the Terminal Arrival area at Taiwan Taoyuan International Airport. Airport taxis provide 24-hour transportation services. For a one way trip from Taoyuan International Airport to Taipei City, the taxi cost falls in the range of NT$1,100 - 1,700 approximately, depending on the numbers of passengers, luggages, and the size of the car.

Terminal 1 Taxi Service Center's phone number: +886-3-398-2832
Terminal 2 Taxi Service Center's phone number: +886-3-398-3599

By Public Transportation

From the airport arrival area, take the airport bus to Zhongxiao Fuxing MRT Station. Then, take the MRT (Taipei Rapid Transit System) brown line, to MRT Taipei Zoo. Then, take the bus (#236, 237, 611, 282, BR11, or BR6) across the street from MRT Taipei Zoo Station to NCCU. Total cost is about NT$200.

From Taiwan Taoyuan International Airport to Hotels

Fullon Hotel Taipei East (福容飯店)

- Bannan Line (Blue) towards Nangang Exhibition Center (Blue):
  - Walk to MRT City Hall Station Exit 3
  - Take bus 912 towards Shenkeng direction from the front of the first floor
- Tamsui Line (Red) towards Xindian Station (Green):
  - Walk to MRT Wanlong Station
  - Take bus 660 towards Shenkeng
- Wenhu Line (Brown):
  - Go to MRT Muzha Station, walk forward
  - Take bus 660 to Wanshunliao Station across the street
- Taiwan Tourist Shuttle:
  - Take MRT Wenhu Line to Muzha Station, get off
  - Transfer to Taipei Bus 795 (Taiwan Tourist Shuttle Muzha-Pingxi Line), and get off at Wanshunliao Station

Just Sleep Taipei NTU (捷思旅)

Take Taiwan Railway (TRA) / Taiwan High-Speed Rail (THSR) to Taipei Station:

- Transfer to the MRT Songshan-Xindian Line and go to Gongguan Station, Exit 2
- Walk for about 2 minutes to reach the hotel

Home Hotel-Xinyi (Home Hotel 信義)

- From Taoyuan Airport to City Hall Bus Station, take the airport bus (Route 1960):
  - Operating hours are from 6:00 AM to 1:00 AM, with buses every 30 minutes
  - Travel time is 60-70 minutes
  - Get off at the Grand Hyatt Hotel stop, then head towards VieShow Cinemas / NEO19
  - Walk for about 10 minutes
- From MRT City Hall Station (Blue Line), Exit 3 to Xinyi Shopping District:
  - The hotel entrance is located behind NEO19
  - Walk for about 10 minutes
- From MRT Xiangshan Station (Red Line), Exit 1 to Xinyi Shopping District:
  - The hotel entrance is located behind NEO19
  - Walk for about 5 minutes

Howard Civil Service International House

(公務人力發展學院福華國際文教會館)

MRT:

- From MRT Taipower Building Station, Exit 2:
  - After exiting, turn left and walk along Xinhai Road for about 10-15 minutes to the intersection of Xinhai Road and Xinsheng South Road
  - Turn left at the intersection to reach the Howard Civil Service International House

GoodMore Hotel Shida (谷墨商旅)

- From Taipei Station, take the Heping Main Line for about 14 minutes to NTNU Union building. After a 2-minute walk, you will reach the Hotel Proverbs Taipei.
- Take the MRT Songshan-Xindian Line (Green Line) to Taipower Building Station. From there, take bus 672, 278, or 949 to NTNU Union building. After a 2-minute walk, you will reach the Hotel Proverbs Taipei.

Guide Hotel Taipei NTU (承攜行旅台北台大館)

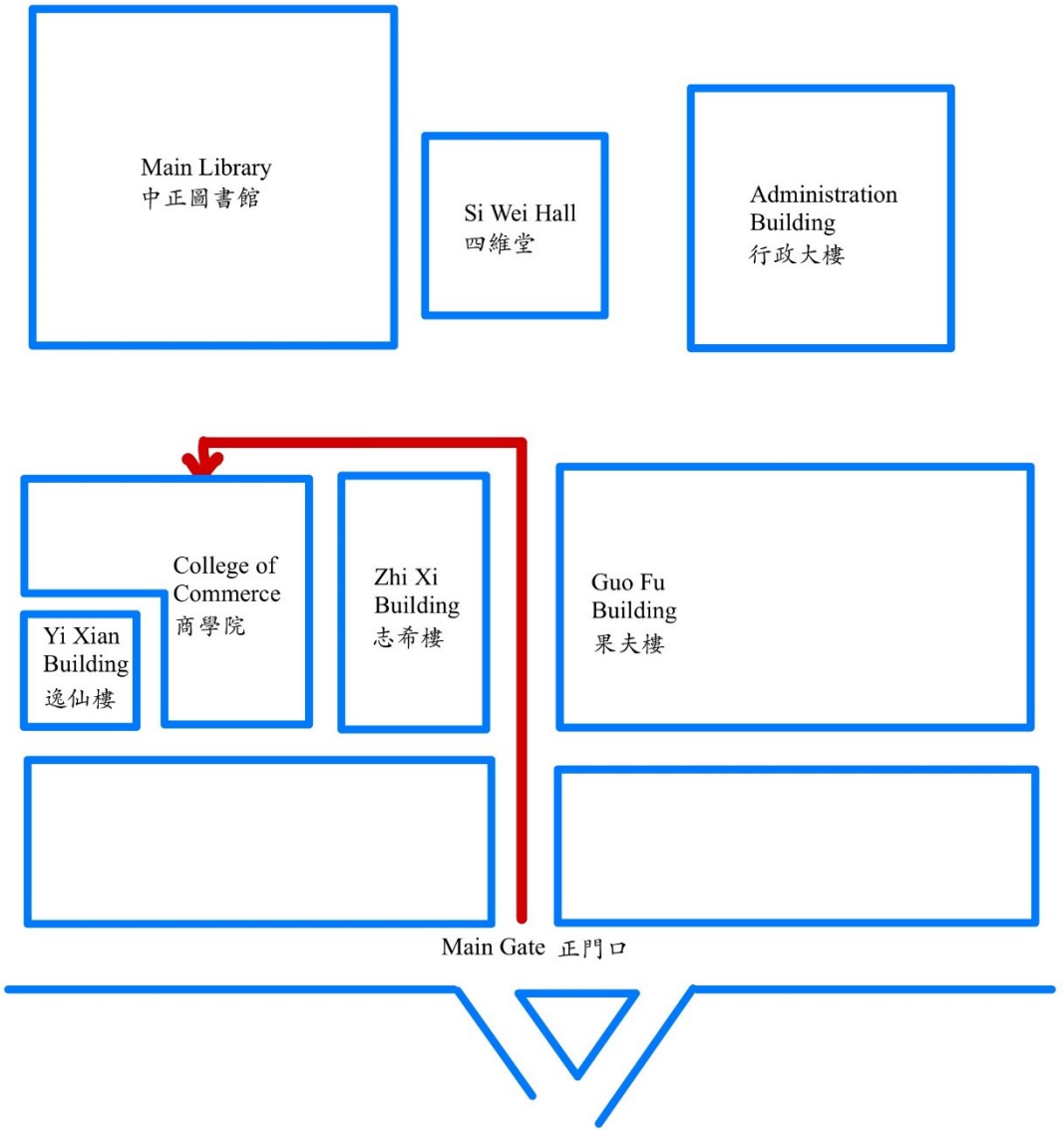
Take the MRT Songshan-Xindian Line (Green Line) to Taipower Building Station. Walk for two minutes to reach your destination.

Banquet information

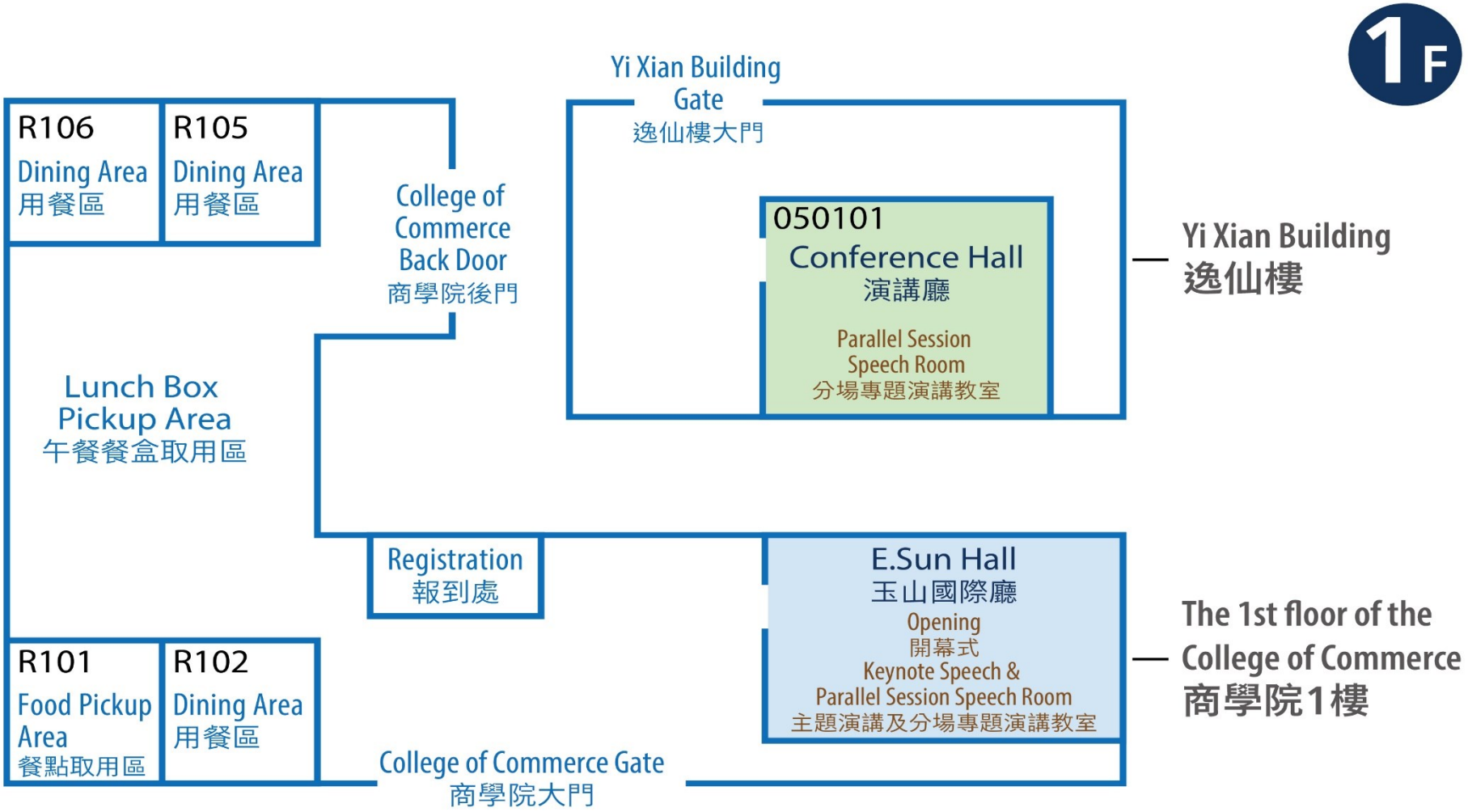
- **Date/time**: 2024/07/09, 17:40
- **Place**: Spring of Shang-Hai Restaurant (春申食府)
  - **Website**: https://springofshanghai.net/
  - **Address**: B1, No. 66, Section 4, Ren'ai Road, Da'an District, Taipei City 106, Taiwan (台北市大安區仁愛路四段 66 號 B1)
  - **Phone**: 02-2707-3555
  - **Transportation Information** (Go there on your own):
    See https://springofshanghai.net/aboutus/

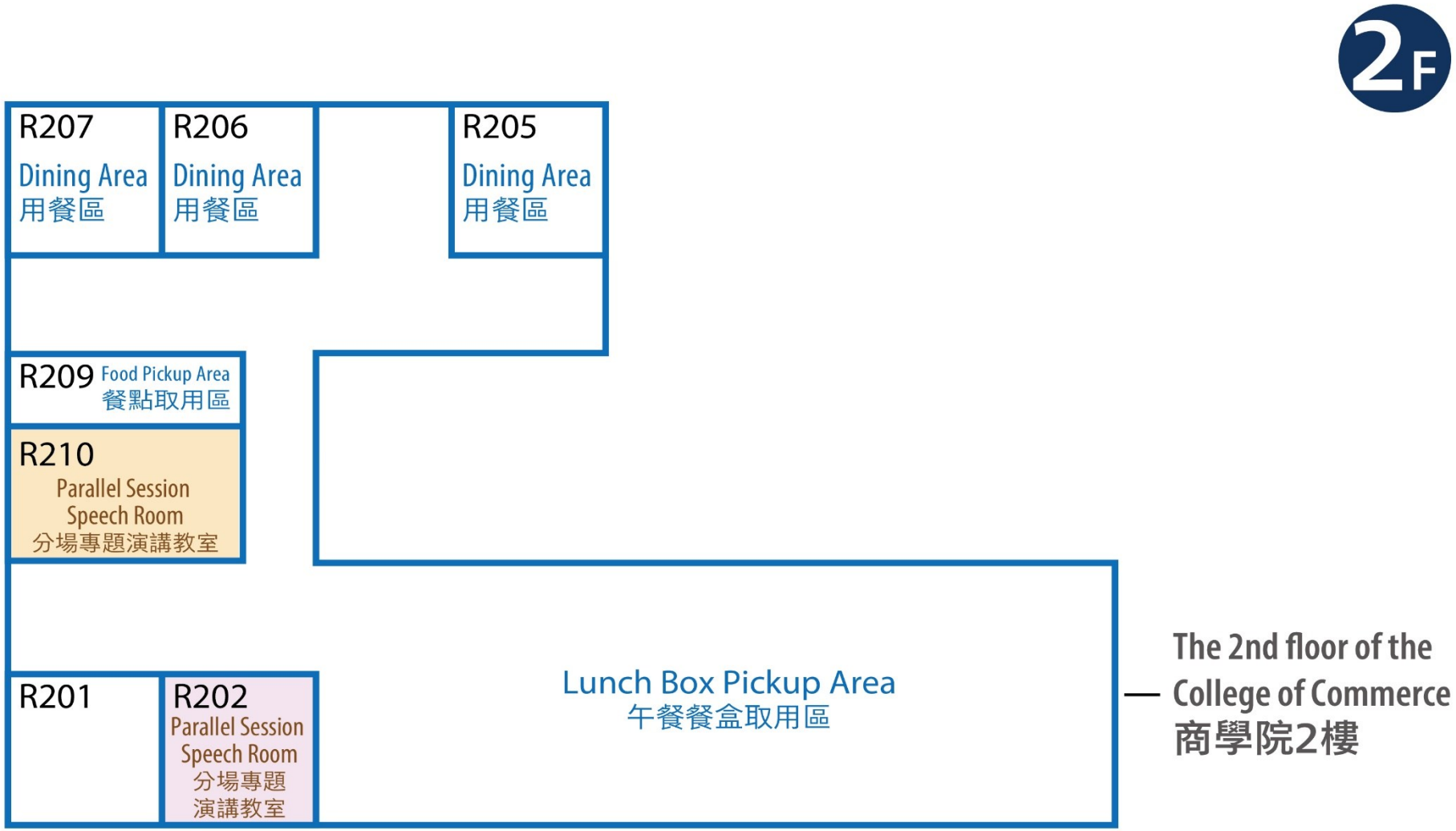
- **Shuttle Bus (NCCU) to Banquet place (one-way trip):**
  - **Gathering Time**: 16:50~17:00
  - **Gathering Location**: Gather at the "check-in" counter on the 1st floor of College of Commerce, and proceed to the boarding area under the guidance of students (National Chengchi University Library).
    (請在商學院一樓的「報到處」集合，將由學生帶隊前往接駁車搭車處(中正圖書館前)。

NCCU Map


Main Library
中正圖書館
Si Wei Hall
四維堂
Administration Building
行政大樓
College of Commerce
商學院
Yi Xian Building
逸仙樓
Zhi Xi Building
志希樓
Guo Fu Building
果夫樓
Main Gate 正門口

Conference Location Map




2F
R207
Dining Area
用餐區
R206
Dining Area
用餐區
R205
Dining Area
用餐區
R209 Food Pickup Area
餐點取用區
R210
Parallel Session
Speech Room
分場專題演講教室
R201
R202
Parallel Session
Speech Room
分場專題
演講教室
Lunch Box Pickup Area
午餐餐盒取用區
The 2nd floor of the
College of Commerce
商學院2樓

Agenda

2024 International Conference for Statistics and Data Science, ICSDS

Date：July 9-10, 2024

Venue：College of Commerce, National Chengchi University

<table>
<tr><td colspan="4">July 9, 2024 (Tue)</td></tr>
<tr><td>Time</td><td colspan="3">Agenda</td></tr>
<tr><td>08:30-
09:00</td><td colspan="3">Guest registration (The 1st floor of College of Commerce)
來賓報到(商學院一樓大廳)
Break (Room 101, the 1st floor of College of Commerce)</td></tr>
<tr><td>09:00-
09:30</td><td colspan="2">Opening Remarks
(E. Sun Hall, the 1st floor of College of Commerce)
開場致詞(商學院一樓玉山國際廳)
Group photo
大合照</td><td>Tsai-Yen Li, President (李蔡彥 校長)
National Chengchi University (國立政治大學)
Jia-Chi Huang, Dean (黃家齊 院長)
College of Commerce, NCCU (國立政治大學商學院)
Su-Fen Yang, Distinguished Professor and Chair
(楊素芬 系主任)
Department of Statistics, NCCU (國立政治大學統計系)</td></tr>
<tr><td>Time</td><td>Agenda</td><td>Speaker</td><td>Chair</td></tr>
<tr><td>9:30-
10:30</td><td>Keynote Speech
(E.Sun Hall, the 1st floor of College of Commerce)
Title：Autoregressive Networks with Stylized Features</td><td>Qiwei Yao (姚琦偉)
Department of Statistics
London School of Economics and Political Science, U.K.</td><td>Cathy W. S. Chen (陳婉淑)
Department of Statistics
Feng Chia University</td></tr>
<tr><td>10:30-
10:50</td><td colspan="3">Break (Room 101, the 1st floor of College of Commerce；Room 209, the 2nd floor of College of Commerce)</td></tr>
</table>

session / venue	Session Speech			
	E. Sun Hall	Room 210, the 2nd floor of College of Commerce (260210)	Room 202, the 2nd floor of College of Commerce (260202)	101 Conference Hall Yi-Xian Building (050101)
10:50-12:20	Parallel Session I			
	I-1 **Topic：Time Series Analysis** **Chair：Nan-Jung Hsu (徐南蓉)** **Institute of Statistics, National Tsing Hua University** 1. Jong-Min Kim, Statistics Discipline, Division of Science and Mathematics, University of Minnesota-Morris, Morris, MN, U.S.A. 2. Ning Ning (寧寧) Department of Statistics, University of Texas A&M, U.S.A. 3. Shih-Feng Huang (黃士峰) Graduate Institute of Statistics, National Central University	**I-2** **Topic：High-Dimensional Data** **Chair：Shwu-Rong Grace Shieh (謝叔蓉)** **Institute of Statistical Sciences, Academia Sinica** 1. Liqun Wang (王力群) Department of Statistics, University of Manitoba, Canada 2. Su-Yun Huang (陳素雲) Institute of Statistical Sciences, Academia Sinica 3. Yu-Bo Wang (王昱博) Department of Mathematical and Statistical Sciences, Clemson University, U.S.A.	**I-3** **Topic：Statistics and Data Science** **Chair：Chao A. Hsiung (熊昭)** **Institute of Population Health Sciences National Health Research Institutes** 1. Jianxin Shi (時建新) Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), U.S.A. 2. Dianliang Deng (鄧殿良) Department of Mathematics & Statistics University of Regina, Canada 3. Li-Hsin Chien (簡立欣) Department of Applied Mathematics, Chung Yuan Christian University	**I-4** **Topic：Machine Learning** **Chair：I-Chen Lee (李宜真)** **Department of Statistics, National Cheng Kung University** 1. W.Y. Wendy Lou Dalla Lana School of Public Health University of Toronto, Canada 2. I-Ming Chiu (邱翊銘) Department of Economics, Rutgers University-Camden, U.S.A. 3. Chih-Hao Chang (張志浩) Department of Statistics, National Chengchi University
12:20-13:30	Luncheon (The 1st & 2nd floor of College of Commerce)			

13:30-15:00	Parallel Session II			
	II-1 **Topic：Biomedical Study** **Chair：Hong-Dar Isaac Wu (吳宏達)** **Department of Applied Mathematics** **National Chung Hsing University** 1. Zhezhen Jin (金哲振) Department of Biostatistics, Columbia University, U.S.A. 2. Hua Zhou (周華) Department of Biostatistics, University of California, Los Angeles, U.S.A. 3. Jia-Han Shih (施嘉翰) Department of Applied Mathematics National Sun Yat-sen University	**II-2** **Topic：Survival Analysis** **Chair：Yi-Ting Hwang (黃怡婷)** **Department of Statistics, National Taipei University** 1. Mei-Ling Ting Lee (丁美齡) Department of Epidemiology and Biostatistics, University of Maryland College Park, U.S.A. 2. Jin Zhou (周瑾) Department of Biostatistics, University of California, Los Angeles, U.S.A. 3. Chung Chang (張中) Department of Applied Mathematics National Sun Yat-sen University	**II-3** **Topic：Time Series Analysis** **Chair：Ie-Bin Lian (連怡斌)** **Department of Mathematics, National Changhua University of Education** 1. William W.S. Wei (魏武雄) Statistics, Operations, and Data Science Temple University, U.S.A. 2. Li-Hsien Sun (孫立憲) Department of Statistics, National Central University 3. Hsin-Chieh Wong (翁新傑) Department of Statistics, National Taipei University	**II-4** **Topic：Biostatistics** **Chair：Chia-Hui Huang (黃佳慧)** **Department of Statistics, National Chengchi University** 1. Naisyin Wang (王乃昕) Department of Statistics, University of Michigan, U.S.A. 2. Li-Shan Huang (黃禮珊) Institute of Statistics, National Tsing Hua University 3. Dongdong Li (李東東) Department of Population Medicine Harvard Medical School, U.S.A.
15:00-15:20	Break (Room 101, the 1st floor of College of Commerce；Room 209, the 2nd floor of College of Commerce)			
15:20-16:50	Parallel Session III			
	III-1 **Topic：Industrial Statistics** **Chair：Sheng-Tsaing Tseng (曾勝滄)** **Institute of Statistics, National Tsing Hua University** 1. Dennis KJ Lin (林共進) Department of Statistics, Purdue University, U.S.A. 2. Tsai-Hung Fan (樊采虹) Graduate Institute of Statistics, National Central University 3. I-Tang Yu (俞一唐) Department of Statistics, Tunghai University	**III-2** **Topic：Causal Inference and Lifetime Data Analysis** **Chair：Tsung-Shan Tsou (鄒宗山)** **Graduate Institute of Statistics** **National Central University** 1. Jialiang Li (栗家量) Department of Statistics & Data Science National University of Singapore 2. Yong Chen (陳勇) Department of Biostatistics, University of Pennsylvania, U.S.A. 3. Hsin-wen Chang (張馨文) Institute of Statistical Sciences, Academia Sinica	**III-3** **Topic：Experimental Designs** **Chair：Mong-Na Lo Huang (羅夢娜)** **Department of Applied Mathematics** **National Sun Yat-sen University** 1. Ming-Hung Kao (高銘宏) School of Mathematical and Statistical Sciences, Arizona State University, U.S.A. 2. Hsiang-Ling Hsu (許湘伶) Institute of Statistics, National University of Kaohsiung 3. Cheng-Yu Sun (孫誠佑) Institute of Statistics, National Tsing Hua University	**III-4** **Topic：Deep Learning and High Dimensional Data analysis** **Chair：Yuan, Juan-Ming (袁淵明)** **Department of Data Science and Big Data Analytics, Providence University** 1. Tai-Been Chen (陳泰賓) Department of Radiological Technology Faculty of Medical Technology Teikyo University, Tokyo, Japan 2. Yen-Lung Tsai (蔡炎龍) Department of Mathematical Sciences National Chengchi University 3. An-Shun Tai (戴安順) Department of Statistics, National Cheng Kung University

16:50- 17:00	shuttle bus@ National Chengchi University Library (Gather at the “check-in” counter on the 1st floor of College of Commerce, and proceed to the boarding area under the guidance of students.) 晚宴接駁車停在中正圖書館前。 欲搭乘接駁車者，請於商學院一樓大廳報到處集合，統一由學生帶至接駁車停車處上車。
17:40	Banquet (the restaurant of spring-shanghai) (春申食府)

July 10, 2024 (Wed)			
Time	Agenda		
08:20-09:00	**Guest registration and pre-meeting communication** (The 1st floor of College of Commerce**)** 來賓報到及會前交流(商學院一樓大廳)		
09:00-09:10	**Opening Remarks** (E. Sun Hall, the 1st floor of College of Commerce**)** 開場致詞(商學院一樓玉山國際廳)	**Jun Zhao**, Executive Director International Chinese Statistical Association, U.S.A. **Guan-Hua Huang**, Professor and Director (黃冠華 所長) Institute of Statistics, National Yang Ming Chiao Tung University (國立陽明交通大學統計研究所)	
Time	Agenda	Speaker	Chair
9:10-10:10	**Keynote Speech** (E. Sun Hall, the 1st floor of College of Commerce**)** Title：Deep Learning for Censored Survival Data	Jane-Ling Wang (王建玲) Department of Statistics University of California, Davis, U.S.A.	Henry Horng-Shing Lu (盧鴻興) Institute of Statistics National Yang Ming Chiao Tung University
10:10-10:30	Break (Room 101, the 1st floor of College of Commerce；Room 209, the 2nd floor of College of Commerce)		

session / venue	Session Speech			
	E. Sun Hall	Room 210, the 2nd Floor of College of Commerce (260210)	Room 202, the 2nd floor of College of Commerce (260202)	101 Conference Hall, Yi-Xian Building (050101)
10:30-12:00	Parallel Session IV			
	IV-1 **Topic：Functional Data Analysis & Dimension Reduction** **Chair：Mei-Hui Guo (郭美惠)** **Department of Applied Mathematics** **National Sun Yat-sen University** 1. Hans-Georg Müller Department of Statistics, University of California, Davis, U.S.A. 2. Ci-Ren Jiang (江其衽) Institute of Statistics and Data Science National Taiwan University 3. Lih-Yuan Deng (鄧利源) Department of Mathematical Sciences University of Memphis, U.S.A.	**IV-2** **Topic：Recent Advances in Risk Analysis** **Chair：Jun Zhao(趙駿)** **International Chinese Statistical Association, U.S.A.** 1. Fabrizio Ruggeri Institute of Applied Mathematics and Information Technology, Italian National Research Council, Italia 2. Kyoji Furukawa Biostatistics Center, Kurume University Japan 3. Rachel Huang (黃瑞卿) Department of Finance, National Central University	**IV-3** **Topic：Cutting-Edge Statistical Modeling Approaches for Multifaceted Data** **Chair：Tsung-I Lin (林宗儀)** **Institute of Statistics, National Chung Hsing University** 1. Victor Hugo Lachos Department of Statistics, University of Connecticut, U.S.A. 2. Mohammad Arashi Department of Mathematical Sciences Ferdowsi University of Mashhad, Iran 3. Chang-Yun Lin (林長瑩) Institute of Statistics, National Chung Hsing University	**IV-4** **Topic：Experimental Designs** **Chair：Ray-Bing Chen (陳瑞彬)** **Department of Statistics, National Cheng Kung University** 1. Weng Kee Wong (王永琪) Department of Biostatistics, University of California, Los Angeles, U.S.A. 2. Qian Helen Li StatsVita, LLC, U.S.A. 3. Ming-Chung Chang (張明中) Institute of Statistical Science, Academia Sinica
12:00-13:00	Luncheon (The 1st & 2nd floor of College of Commerce)			

13:00-14:30	Parallel Session V			
	V-1 **Topic：Statistics and Data Science** **Chair：Chien-Tai Lin (林千代)** **Department of Mathematics, Tamkang University** 1. Samuel Kou (寇星昌) Department of Statistics, University of Harvard, U.S.A. 2. Jung-Ying Tzeng (曾仲瑩) Department of Statistics and Bioinformatics Research Center, North Carolina State University, U.S.A. 3. Hsiuying Wang (王秀瑛) Institute of Statistics, National Yang Ming Chiao Tung University	**V-2** **Topic：Deep Learning and AI** **Chair：Chun-Shu Chen (陳春樹)** **Institute of Statistics, National Central Universityy** 1. Fushing Hsieh (謝復興) Department of Statistics, University of California, Davis, U.S.A. 2. Ting-Li Chen (陳定立) Institute of Statistical Sciences, Academia Sinica 3. Guan-Hua Huang (黃冠華) Institute of Statistics, National Yang Ming Chiao Tung University	**V-3** **Topic：Statistics and Data Science** **Chair：Henghsiu Tsai (蔡恆修)** **Institute of Statistical, Sciences Academia Sinica** 1. Ying Hung (洪瑛) Department of Statistics, Rutgers, the State University of New Jersey, U.S.A. 2. Takeshi Emura (江村剛志) Research Center for Medical and Health Data Science, the Institute of Statistical Mathematics, Tokyo, Japan 3. Jeng-Huei Chen (陳政輝) Department of Mathematical Sciences National Chengchi University	**V-4** **Topic：Biostatistics** **Chair：Wen-Han Hwang (黃文瀚)** **Institute of Statistics, National Tsing Hua University** 1. George C. Tseng (曾建城) Department of Biostatistics, University of Pittsburgh, U.S.A. 2. Feng-Chang Lin (林逢章) Department of Biostatistics, University of North Carolina at Chapel Hill, U.S.A. 3. Ming-Yueh Huang (黃名鉞) Institute of Statistical Sciences, Academia Sinica
14:30-14:50	Break (Room 101, the 1st floor of College of Commerce；Room 209, the 2nd floor of College of Commerce)			
14:50-16:20	Parallel Session VI			
	VI-1 **Topic：Biostatistics** **Chair：Ly-Yu D Liu (劉力瑜)** **Department of Agronomy, National Taiwan University** 1. Naitee Ting (丁迺迪) Biostatistics and Data Sciences, Boehringer Ingelheim Pharmaceuticals, Inc. , U.S.A. 2. Wan Yuo Guo (郭萬祐) Taipei Veterans General Hospital and School of Medicine, National Yang Ming Chiao Tung University 3. Il Do Ha Department of Statistics & data Science Pukyong National University South Korea	**VI-2** **Topic：Statistics and Data Science** **Chair：Nan-Cheng Su (蘇南誠)** **Department of Statistics, National Taipei University** 1. Andrei Volodin Department of Mathematics & Statistics University of Regina, Canada 2. Xiao Wang (王嘯) Department of Statistics, College of Science Purdue University, U.S.A. 3. Pei-Ting Chou (周珮婷) Department of Statistics, National Chengchi University	**VI-3** **Topic： Statistical Process Monitoring** **Chair：Chien-Yu Peng (彭健育)** **Institute of Statistical Sciences Academia Sinica** 1. Arthur B. Yeh (葉百堯) Department of Applied Statistics and Operations Research, Bowling Green State University, U.S.A. 2. Wei-Hang Huang (黃偉恆) Department of Statistics, Feng Chia University 3. Ming-Che Lu (呂明哲) Department of Accounting, Chaoyang University of Technology	**VI-4** **Topic：Modern Statistics and Applications** **Chair：Tsai, Pi-Wen (蔡碧紋)** **Department of Mathematics, National Taiwan Normal University** 1.Chih-Li Sung (宋治立) Department of Statistics and Probability Michigan State University, U.S.A. 2.Frederick Kin Hing Phoa (潘建興) Institute of Statistical Science, Academia Sinica 3. Yu-Wei Chang (張育瑋) Department of Statistics, National Chengchi University

Keynote Speeches

Keynote Speech I

July 9, 2024 at 9:30-10:30

E. Sun Hall

- **Chair:**

  Cathy W. S. Chen (陳婉淑), Department of Statistics, Feng Chia University

- **Speaker:**

  Qiwei Yao (姚琦偉), Department of Statistics, London School of Economics and Political Science, U.K.

Autoregressive Networks with Stylized Features

Qiwei Yao

Department of Statistics, London School of Economics and Political Science, U.K.

q.yao@lse.ac.uk

Abstract

We propose a first-order autoregressive model for dynamic network processes in which edges change over time while nodes remain unchanged. The model depicts the dynamic changes explicitly. It also facilitates simple and efficient statistical inference such as the maximum likelihood estimators which are proved to be (uniformly) consistent and asymptotically normal. The model diagnostic checking can be carried out easily using a permutation test. We also elucidate how the AR model can accommodate node heterogeneity, edge sparsity, transitivity, homophily and other stylized features in network data.

Keynote Speech II

July 10, 2024 at 9:10-10:10

E. Sun Hall

◆ **Chair:**

Henry Horng-Shing Lu (盧鴻興), Institute of Statistics, National Yang Ming Chiao Tung University

◆ **Speaker:**

Jane-Ling Wang (王建玲), Department of Statistics, University of California, Davis, U.S.A.

Deep Learning for Censored Survival Data

Jane-Ling Wang

Department of Statistics, University of California, Davis, U.S.A.

janelwang@ucdavis.edu

Abstract

Unlike standard tasks, survival analysis requires modeling incomplete data, such as right-censored data, which must be treated with care. While deep neural networks excel in traditional supervised learning, it remains unclear how to best utilize these models in survival analysis. A key question asks which data-generating assumptions of traditional survival models should be retained and which should be made more flexible via the function-approximating capabilities of neural networks. In addition, most of these methods are difficult to interpret and mathematical understanding of them is lacking. In this talk, we explore these issues from two directions. First, we study the partially linear Cox model, where the nonlinear component of the model is implemented using a deep neural network. The proposed approach is flexible and able to circumvent the curse of dimensionality, yet it facilitates interpretability of the effects of treatment covariates on survival. Next, we introduce a Deep Extended Hazard (DeepEH) model to provide a flexible and general framework for deep survival analysis. The extended hazard model includes the conventional Cox proportional hazards and accelerated failure time models as special cases, so DeepEH subsumes the popular Deep Cox proportional hazard (DeepSurv) and Deep Accelerated Failure Time (DeepAFT) models. We provide theoretical support for the proposed models, which underscores the attractive feature that deep learning is able to detect low-dimensional structure of data in high-dimensional space. Numerical experiments further provide evidence that the proposed methods outperform existing statistical and deep learning approaches to survival analysis. Time permitting, we will explore how to perform hypothesis testing for survival data.

Session Agenda

Parallel Session I schedule

July 9, 2024 at 10:50-12:20

I-1 Topic: Time Series Analysis

E. Sun Hall

◆ **Chair:**

Nan-Jung Hsu (徐南蓉), Institute of Statistics, National Tsing Hua University

◆ **Speakers:**

1. Jong-Min Kim, Statistics Discipline, Division of Science and Mathematics, University of Minnesota-Morris, Morris, MN, U.S.A.
   Change Point Detection for the Intraday Volatility Using Functional ARCH and Conditional Copula
2. Ning Ning (寧寧), Department of Statistics, University of Texas A&M, U.S.A.
   Advancements in Online Learning for High-Dimensional Spatiotemporal Models
3. Shih-Feng Huang (黃士峰), Graduate Institute of Statistics, National Central University
   Hysteretic Multivariate Bayesian Structural GARCH Model with Soft Information

I-2 Topic: High-Dimensional Data

Room 260210

◆ **Chair:**

Shwu-Rong Grace Shieh (謝叔蓉), Institute of Statistical Sciences, Academia Sinica

◆ **Speakers:**

1. Liqun Wang (王力群), Department of Statistics, University of Manitoba, Canada
   Regularized Estimation of Covariance Matrix in High-Dimensional Models
2. Su-Yun Huang (陳素雲), Institute of Statistical Sciences, Academia Sinica
   A Geometric Algorithm for Contrastive PCA in High Dimension
3. Yu-Bo Wang (王昱博), Department of Mathematical and Statistical Sciences, Clemson University, U.S.A.
   SIGHR: Side Information Guided High-Dimensional Regression

I-3 Topic: Statistics and Data Science

Room 260202

◆ **Chair:**

Chao A. Hsiung (熊昭), Institute of Population Health Sciences, National Health Research Institutes

◆ **Speakers:**

1. Jianxin Shi (時建新), Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), U.S.A.
   Estimating Overall Fraction of Phenotypic Variance Attributed to Predictors Measured with Error
2. Dianliang Deng (鄧殿良), Department of Mathematics & Statistics, University of Regina, Canada
   Adaptive Aggregation for Longitudinal Quantile Regression Based on Censored History Process
3. Li-Hsin Chien (簡立欣), Department of Applied Mathematics, Chung Yuan Christian University
   Developing and Validating Absolute Risk Prediction Models for Breast Cancer in Taiwan with Synthesized Data from Multiple Sources

I-4 Topic: Machine Learning

050101 Conference Hall

◆ **Chair:**

I-Chen Lee (李宜真), Department of Statistics, National Cheng Kung University

◆ **Speakers:**

1. W.Y. Wendy Lou, Dalla Lana School of Public Health, University of Toronto, Canada
   Fusion Clustering for Multi-Source Longitudinal Data

2. I-Ming Chiu (邱翊銘), Department of Economics, Rutgers University-Camden, U.S.A.
   Predicting Adolescent Depression in the U.S. Using a Two-Layered Ensemble Machine Learning Approach

3. Chih-Hao Chang (張志浩), Department of Statistics, National Chengchi University
   Application of Machine Learning Techniques in Threshold Regression Modeling

Parallel Session II schedule

July 9, 2024 at 13:30-15:00

II-1 Topic: Biomedical Study

E. Sun Hall

◆ **Chair:**

Hong-Dar Isaac Wu (吳宏達) Department of Applied Mathematics, National Chung Hsing University

◆ **Speakers:**

1. Zhezhen Jin (金哲振), Department of Biostatistics, Columbia University, U.S.A.

   Semiparametric Statistical Methods for the Analysis of Biomedical Data

2. Hua Zhou (周華), Department of Biostatistics, University of California, Los Angeles, U.S.A.

   Scalable and Robust Censored Linear Regression with Applications to Biobank Studies

3. Jia-Han Shih (施嘉翰), Department of Applied Mathematics, National Sun Yat-sen University

   A Class of Regression Association Measure Based on Concordance

II-2 Topic: Survival Analysis

Room 260210

◆ **Chair:**

Yi-Ting Hwang (黃怡婷), Department of Statistics, National Taipei University

◆ **Speakers:**

1. Mei-Ling Ting Lee (丁美齡), Department of Epidemiology and Biostatistics, University of Maryland, College Park, U.S.A.

   Neural Network Extension of Threshold Regression for Event-Time Data

2. Jin Zhou (周瑾), Department of Biostatistics, University of California, Los Angeles, U.S.A.

   Scalable and Robust Joint Models for Longitudinal and Survival Outcomes

3. Chung Chang (張中), Department of Applied Mathematics, National Sun Yat-sen University

   Heavy-Tailed Distribution for Combining Dependent P-Values with Asymptotic Robustness

II-3 Topic: Time Series Analysis.

Room 260202

◆ **Chair:**

Ie-Bin Lian (連怡斌), Department of Mathematics, National Changhua University of Education

◆ **Speakers:**

1. William W.S. Wei (魏武雄), Statistics, Operations, and Data Science, Temple University, U.S.A.

   Issues on Time Series Analysis and Applications

2. Li-Hsien Sun (孫立憲), Department of Statistics, National Central University

   Interval-Based Time Series Analysis: Detecting Structural Shifts and Change-Point Estimation

3. Hsin-Chieh Wong (翁新傑), Department of Statistics, National Taipei University

   Valid Post-Averaging Inference in AR-G/GARCH Models

II-4 Topic: Biostatistics

050101 Conference Hall

◆ **Chair:**

Chia-Hui Huang (黃佳慧), Department of Statistics, National Chengchi University

◆ **Speakers:**

1. Naisyin Wang (王乃昕), Department of Statistics, University of Michigan, U.S.A.

   Utilizing Synthetic Components to Balance Privacy Protection and Data Utility

2. Li-Shan Huang (黃禮珊), Institute of Statistics, National Tsing Hua University

   Promotion Time Cure Model with Local Polynomial Estimation

3. Dongdong Li (李東東), Department of Population Medicine, Harvard Medical School, U.S.A.

   Evaluating Association Between Two Event Times with Observations Subject to Informative Censoring

Parallel Session III schedule

July 9, 2024 at 15:20-16:50

III-1 Topic: Industrial Statistics

E. Sun Hall

◆ **Chair:**

Sheng-Tsaing Tseng (曾勝滄), Institute of Statistics, National Tsing Hua University

◆ **Speakers:**

1. Dennis KJ Lin (林共進), Department of Statistics, Purdue University, U.S.A.

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

2. Tsai-Hung Fan (樊采虹), Graduate Institute of Statistics, National Central University

   General Random-Effects Trend Renewal Processes with Applications

3. I-Tang Yu (俞一唐), Department of Statistics, Tunghai University

   Increment Degradation Model: A Bayesian Perspective

III-2 Topic: Causal Inference and Lifetime Data Analysis

Room 260210

◆ **Chair:**

Tsung-Shan Tsou (鄒宗山), Graduate Institute of Statistics, National Central University

◆ **Speakers:**

1. Jialiang Li (栗家量), Department of Statistics & Data Science, National University of Singapore

   Efficient Auxiliary Information Synthesis for Cure Rate Model

2. Yong Chen (陳勇), Department of Biostatistics, University of Pennsylvania, U.S.A.

   Real-World Effectiveness of BNT162b2 Against Infection in Children: Causal Inference Under Misclassification in Treatment Status

3. Hsin-wen Chang (張馨文), Institute of Statistical Sciences, Academia Sinica

   Bivariate Analysis of Distribution Functions Under Biased Sampling

III-3 Topic: Experimental Designs

Room 260202

◆ **Chair:**

Mong-Na Lo Huang (羅夢娜), Department of Applied Mathematics, National Sun Yat-sen University

◆ **Speakers:**

1. Ming-Hung Kao (高銘宏), School of Mathematical and Statistical Sciences, Arizona State University, U.S.A.
   Optimal Study Designs for Sparse Functional Data Analysis
2. Hsiang-Ling Hsu (許湘伶), Institute of Statistics, National University of Kaohsiung
   Optimal Designs with Multiple Correlated Responses for Experiments with Mixtures
3. Cheng-Yu Sun (孫誠佑), Institute of Statistics, National Tsing Hua University
   Space-Filling Regular Designs Under a Minimum Aberration-Type Criterion

III-4 Topic: Deep Learning and High Dimensional Data Analysis

050101 Conference Hall

◆ **Chair:**

Yuan, Juan-Ming (袁淵明), Department of Data Science and Big Data Analytics, Providence University

◆ **Speakers:**

1. Tai-Been Chen (陳泰賓), Department of Radiological Technology, Faculty of Medical Technology, Teikyo University, Tokyo, Japan
   Deep Learning Applications in Chest X-Ray Classification, CT Liver Tumors Segmentation, and Detection of Stenosis on X-Ray Coronary Angiography
2. Yen-Lung Tsai (蔡炎龍), Department of Mathematical Sciences, National Chengchi University
   Contrastive Learning for Time Series Data: Predicting Stock Price Movements
3. An-Shun Tai (戴安順), Department of Statistics, National Cheng Kung University
   Robust and Flexible High-Dimensional Causal Mediation Model for DNA Methylation Studies

Parallel Session IV schedule

July 10, 2024 at 10:30-12:00

IV-1 Topic: Functional Data Analysis & Dimensional Reduction.

E. Sun Hall

◆ **Chair:**

Mei-Hui Guo (郭美惠), Department of Applied Mathematics, National Sun Yat-sen University

◆ **Speakers:**

1. Hans-Georg Müller – Department of Statistics, University of California, Davis, U.S.A.
   Quantifying Variation for Random Objects Via Distance Profiles
2. Ci-Ren Jiang (江其衽), Institute of Statistics and Data Science, National Taiwan University
   Eigen-Adjusted FPCA
3. Lih-Yuan Deng (鄧利源) – Department of Mathematical Sciences, University of Memphis, U.S.A.
   Big Data Model Building Using Dimension Reduction and Sample Selection

IV-2 Topic: Recent Advances in Risk Analysis

Room 260210

◆ **Chair:**

Jun Zhao, International Chinese Statistical Association, U.S.A.

◆ **Speakers:**

1. Fabrizio Ruggeri – Institute of Applied Mathematics and Information Technology, Italian National Research Council, Italia.
   Recent Advances in Adversarial Risk Analysis
2. Kyoji Furukawa – Biostatistics Center, Kurume University, Japan
   Statistical Challenges in Radiation Risk Assessment
3. Rachel Huang (黃瑞卿), Department of Finance, National Central University
   Hedge Funds Performance: Are Crypto Hedge Funds the Rising Star?

IV-3 Topic: Cutting-Edge Statistical Modeling Approaches for Multifaceted Data

Room 260202

◆ **Chair:**

Tsung-I Lin (林宗儀), Institute of Statistics, National Chung Hsing University

◆ **Speakers:**

1. Victor Hugo Lachos – Department of Statistics, University of Connecticut, U.S.A.
   Heckman Selection Contaminated Normal Model: Parameter Estimation via the EM-Algorithm
2. Mohammad Arashi – Department of Mathematical Sciences, Ferdowsi University of Mashhad, IRAN
   High-dimensional Regression Analysis with Machine Learning
3. Chang-Yun Lin (林長鋆), Institute of Statistics, National Chung Hsing University
   Design Construction and Model Selection for Small Mixture-Process Variable (MPV) Experiments with High-Dimensional Model Terms

IV-4 Topic: Experimental Design

050101 Conference Hall

◆ **Chair:**

Ray-Bing Chen (陳瑞彬), Department of Statistics, National Cheng Kung University

◆ **Speakers:**

1. Weng Kee Wong (王永琪) – Department of Biostatistics, University of California, Los Angeles, U.S.A.
   Optimal Exact Designs for Small Studies in Toxicology with Applications to Hormesis via Metaheuristics
2. Qian Helen Li, StatsVita, LLC, U.S.A.
   Extending the Use of Adaptive Design Beyond Sample Size Re-Estimation
3. Ming-Chung Chang (張明中), Institute of Statistical Science, Academia Sinica
   Orthogonalized Moment Aberration for Multi-Stratum Factorial Designs

Parallel Session V schedule

July 10, 2024 at 13:00-14:30

V-1 Topic: Statistics and Data Science

E. Sun Hall

◆ **Chair:**

Chien-Tai Lin (林千代), Department of Mathematics, Tamkang University

◆ **Speakers:**

1. Samuel Kou (寇星昌), Department of Statistics, University of Harvard, U.S.A.

   Statistical Inference of Dynamic Systems via Manifold-Constrained Gaussian Processes

2. Jung-Ying Tzeng (曾仲瑩), Department of Statistics and Bioinformatics Research Center, North Carolina State University, U.S.A.

   Transfer Learning with False Negative Control Improves Polygenic Risk Prediction

3. Hsiuying Wang (王秀瑛), Institute of Statistics, National Yang Ming Chiao Tung University

   Tolerance Interval for COVID-19 Data Prediction

V-2 Topic: Deep Learning + AI

Room 260210

◆ **Chair:**

Chun-Shu Chen (陳春樹), Institute of Statistics, National Central University

◆ **Speakers:**

1. Fushing Hsieh (謝復興), Department of Statistics, University of California, Davis, U.S.A.

   Data Intelligence vs A.I.

2. Ting-Li Chen (陳定立), Institute of Statistical Sciences, Academia Sinica

   Multiscale Major Factor Selections for Complex System Data with Structural Dependency and Heterogeneity

3. Guan-Hua Huang (黃冠華), Institute of Statistics, National Yang Ming Chiao Tung University

   Automated Convolutional Neural Network and Transformer for Multi-Class Classification of Three-Dimensional Brain Images

V-3 Topic: Statistics and Data Science

Room 260202

◆ **Chair:**

Henghsiu Tsai (蔡恆修), Institute of Statistical Sciences, Academia Sinica

◆ **Speakers:**

1. Ying Hung (洪瑛), Department of Statistics, Rutgers, the State University of New Jersey, U.S.A.

   Analysis and Uncertainty Quantification of Digital Twins

2. Takeshi Emura (江村剛志), Research Center for Medical and Health Data Science, the Institute of Statistical Mathematics, Tokyo, Japan

   Survival Prognostic Analysis with Copula-Graphic Estimator for Dependent Censoring

3. Jeng-Huei Chen (陳政輝), Department of Mathematical Sciences, National Chengchi University

   A Big-Data Based Model of Chronic Kidney Disease and Its Applications

V-4 Topic: Biostatistics

050101 Conference Hall

◆ **Chair:**

Wen-Han Hwang (黃文瀚), Institute of Statistics, National Tsing Hua University

◆ **Speakers:**

1. George C. Tseng (曾建城), Department of Biostatistics, University of Pittsburgh, U.S.A.

   Outcome-Guided Disease Subtyping by Generative Model and Weighted Joint Likelihood in Omics Applications

2. Feng-Chang Lin (林逢章), Department of Biostatistics, University of North Carolina at Chapel Hill, U.S.A.

   Maximum Likelihood Estimation of the Silent Hypnozoite Carriage in a Malaria Randomized Clinical Trial

3. Ming-Yueh Huang (黃名鉞), Institute of Statistical Sciences, Academia Sinica

   Improved Estimation Under Proportional Rates Model for Recurrent Events Data

Parallel Session VI schedule

July 10, 2024 at 14:50-16:20

VI-1 Topic: Biostatistics

E. Sun Hall

◆ **Chair:**

Ly-Yu D Liu (劉力瑜), Department of Agronomy, National Taiwan University

◆ **Speakers:**

1. Naitee Ting (丁迺迪), Biostatistics and Data Sciences, Boehringer Ingelheim Pharmaceuticals, Inc., U.S.A.

   New Drug Development and Dose Finding

2. Wan Yuo Guo (郭萬祐), Taipei Veterans General Hospital and School of Medicine, National Yang Ming Chiao Tung University

   The Journey of Imaging AI: From Data Governance to Clinical Implementation

3. Il Do Ha, Department of Statistics & data Science, Pukyong National University, Busan, South Korea

   Deep Neural Networks for Frailty Models with Clustered Survival Data

VI-2 Topic: Statistics and Data Science

Room 260210

◆ **Chair:**

Nan-Cheng Su (蘇南誠), Department of Statistics, National Taipei University

◆ **Speakers:**

1. Andrei Volodin, Department of Mathematics & Statistics, University of Regina, Canada

   Statistical Inference for the Ratio of Medians of Two Lognormal Distributions

2. Xiao Wang (王嘯), Department of Statistics, College of Science, Purdue University, U.S.A.

   Efficient Multi-modal Sampling via Tempered Distribution Flow

3. Pei-Ting Chou (周珮婷), Department of Statistics, National Chengchi University

   Unveiling the Power of Dimension Reduction in Neural Networks

VI-3 Topic: Statistical Process Monitoring

Room 260202

◆ **Chair:**

Chien-Yu Peng (彭健育), Institute of Statistical Sciences, Academia Sinica

◆ **Speakers:**

1. Arthur B. Yeh (葉百堯), Department of Applied Statistics and Operations Research, Bowling Green State University, U.S.A.

   On Monitoring and Post-Detection Diagnostics of Correlated Quality Variables of Different Types

2. Wei-Hang Huang (黃偉恆), Department of Statistics, Feng Chia University

   The Performance of S Control Charts for the Lognormal Distribution with Estimated Parameters

3. Ming-Che Lu (呂明哲), Department of Accounting, Chaoyang University of Technology

   Pollution Concentration Monitoring Using a New Birnbaum-Saunders Control Chart

VI-4 Topic: Modern Statistics and Applications

050101 Conference Hall

◆ **Chair:**

Tsai, Pi-Wen (蔡碧紋), Department of Mathematics, National Taiwan Normal University

◆ **Speakers:**

1. Chih-Li Sung (宋治立), Department of Statistics and Probability, Michigan State University, U.S.A.

   Stacking Designs: Designing Multi-Fidelity Computer Experiments with Target Predictive Accuracy

2. Frederick Kin Hing Phoa (潘建興), Institute of Statistical Science, Academia Sinica

   An Efficient Approach for Identifying Important Biomarkers for Biomedical Diagnosis

3. Yu-Wei Chang (張育瑋), Department of Statistics, National Chengchi University

   Bayesian Inference with Spike-and-Slab Priors for Differential Item Functioning Detection in a Multiple-Group Irt Tree Model

Abstract

Parallel Session I

Session I-1: Time Series Analysis
July 9, 2024 at 10:50-12:20
E. Sun Hall

◆ **Chair:**

Nan-Jung Hsu (徐南蓉), Institute of Statistics, National Tsing Hua University

◆ **Speakers:**

1. Jong-Min Kim, Statistics Discipline, Division of Science and Mathematics, University of Minnesota-Morris, Morris, MN, U.S.A.
2. Ning Ning (寧寧), Department of Statistics, University of Texas A&M, U.S.A.
3. Shih-Feng Huang (黃士峰), Graduate Institute of Statistics, National Central University

Change Point Detection for the Intraday Volatility Using Functional ARCH and Conditional Copula

Jong-Min Kim

Statistics Discipline, Division of Science and Mathematics, University of Minnesota-Morris, Morris, MN, USA

jongmink@morris.umn.edu

Abstract

In this research, we are concerned with intraday volatilities computed by functional ARCH(1) (fARCH(1), for short) model for high-frequency financial time series. A conditional-Copula multiple change point detection (CPD) for intraday volatilities is proposed using fARCH(1), bivariate Gaussian Copula and t-Copula conditional distributions. We employ current available multivariate CPD models which include energy test based control chart (ETCC) and nonparametric multivariate change point model (NPMVCP) to implement the proposed CPD method for the intraday volatilities. A simulation study is conducted to demonstrate that the functional ARCH based conditional-Copula CPD for the intraday volatilities can be a useful econometrics method to detect abnormal intraday volatilities in the financial market. We analyze intraday volatilities of the Korea composite stock price index (KOSPI) and the Hyundai-Motor (HDM) company stock data with one minute high-frequency to illustrate our proposed CPD method.

Advancements in Online Learning for High-Dimensional Spatiotemporal Models

Ning Ning

Department of Statistics, University of Texas A&M, U.S.A.

patning@tamu.edu

Abstract

Over the past decade, there has been a notable surge in the availability of extensive datasets characterized by their substantial scale. General spatiotemporal dynamical models have emerged as valuable tools for handling data exhibiting both temporal and spatial dependencies. In the realm of complex spatiotemporal systems, where both joint and marginal distributions are often unknown, recent advancements in machine learning have been motivated by the challenges posed by large-scale data. To overcome the high-dimensional spatiotemporal challenge, a successful machine learning algorithm should possess qualities such as applicability, interpretability, explainability, efficiency, and the ability to mitigate the curse of dimensionality. In this presentation, I will delve into an online learning algorithm designed to meet these criteria, allowing for the inference of high-dimensional latent states and unknown parameters across general spatiotemporal models.

Hysteretic Multivariate Bayesian Structural GARCH Model with Soft Information

Shih-Feng Huang

Graduate Institute of Statistics National Central University

huangsf525@gmail.com

Abstract

This study proposes a hysteretic multivariate Bayesian structural GARCH model with soft information, denoted by SH-MBS-GARCH, to describe multidimensional financial time-series dynamics. We first filter the GARCH effects inherent in each financial time series by the De-GARCH technique. Next, we establish a hysteretic multivariate Bayesian structural model for the multidimensional De-GARCH time series to simultaneously capture the trend, seasonal, cyclic, and endogenous (or exogenous) covariates' effects. In particular, we extract soft information from the daily financial news and add the information into the hysteretic part of the model to reflect economic effects on the time-series behavior. An MCMC algorithm is proposed for parameter estimation. The empirical study employs the Dow Jones Industrial, Nasdaq, and Philadelphia Semiconductor indices from January 2016 to December 2020 to investigate the performances of the proposed model. Numerical results reveal that the SH-MBS-GARCH model has better fitting and prediction performances than competitors.

Session I-2: High-Dimensional Data
July 9, 2024 at 10:50-12:20
Room 210, the 2nd floor of College of Commerce (260210)

◆ **Chair:**

Shwu-Rong Grace Shieh (謝叔蓉), Institute of Statistical Sciences, Academia Sinica

◆ **Speakers:**

1. Liqun Wang (王力群), Department of Statistics, University of Manitoba, Canada
2. Su-Yun Huang (陳素雲), Institute of Statistical Sciences, Academia Sinica
3. Yu-Bo Wang (王昱博), Department of Mathematical and Statistical Sciences, Clemson University, U.S.A.

Regularized Estimation of Covariance Matrix in High-Dimensional Models

Liqun Wang

Department of Statistics, University of Manitoba, Canada

liqun.wang@umanitoba.ca

Abstract

Estimation of high-dimensional covariance matrix is one of the fundamental and important problems in multivariate analysis and has a wide range of applications in many fields. This paper presents a novel method for sparse covariance matrix estimation via solving a non-convex regularization optimization problem. We establish the asymptotic properties of the proposed estimator and develop a multi-stage convex relaxation method that guarantees any accumulation point of the sequence generated is a first-order stationary point of the non-convex optimization. Moreover, the error bounds of the first two stage estimators of the multi-stage convex relaxation method are derived under some regularity conditions. Numerical results show that our estimator outperforms the state-of-the-art estimators and has a high degree of sparsity.

A Geometric Algorithm for Contrastive PCA in High Dimension

Su-Yun Huang

Institute of Statistical Sciences, Academia Sinica

syhuang@stat.sinica.edu.tw

Abstract

Principal component analysis (PCA) has been widely used in exploratory data analysis. Contrastive PCA (Abid et al.), a generalized method of PCA, is a new tool used to capture features of a target dataset relative to a background dataset while preserving the maximum amount of information contained in the data. With high dimensional data, contrastive PCA becomes impractical due to its high computational requirement of forming the contrastive covariance matrix and associated eigenvalue decomposition for extracting leading components. In this article, we propose a geometric curvilinear-search method to solve this problem and provide a convergence analysis. Our approach offers significant computational efficiencies. Specifically, it reduces the time complexity from O((n ∨ m)p2) to a more manageable O((n ∨ m)pr), where n, m are the sample sizes of the target data and background data, respectively, p is the data dimension and r is the number of leading components. Additionally, we streamline the space complexity from O(p2), necessary for storing the contrastive covariance matrix, to a more economical O((n ∨ m)p), sufficient for storing the data alone. Numerical examples are presented to show the merits of the proposed algorithm.

SIGHR: Side Information Guided High-Dimensional Regression

Yu-Bo Wang

Department of Mathematical and Statistical Sciences, Clemson University, U.S.A.

yubow@clemson.edu

Abstract

In this work, we develop a novel Bayesian regression framework that can be used to complete variable selection in high dimensional settings. Unlike existing techniques, the proposed approach can leverage auxiliary information to inform about the sparsity structure of regression coefficients. This is accomplished by replacing the usual inclusion probability in the spike and slab prior with a binary regression model which assimilates this extra source of information. To facilitate model fitting, a computationally efficient and easy to implement MCMC posterior sampling algorithm is developed via carefully chosen priors and data augmentation steps. The finite sample performance of our methodology is assessed through numerical simulations, and we further illustrate our approach by using it to identify genetic markers associated with nicotine metabolism rate; a key biological marker associated with nicotine dependence.

Keywords: Side information; High dimensional regression; Spike and slab prior; Conditional means prior; Nicotine metabolite ratio.

Session I-3: Statistics and Data Science
July 9, 2024 at 10:50-12:20
Room 202, the 2nd floor of College of Commerce (260202)

◆ **Chair:**

Chao A. Hsiung (熊昭), Institute of Population Health Sciences, National Health Research Institutes

◆ **Speakers:**

1. Jianxin Shi (時建新), Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), U.S.A.
2. Dianliang Deng (鄧殿良), Department of Mathematics & Statistics, University of Regina, Canada
3. Li-Hsin Chien (簡立欣), Department of Applied Mathematics, Chung Yuan Christian University

Estimating Overall Fraction of Phenotypic Variance Attributed to Predictors Measured with Error

Jianxin Shi

Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), U.S.A.

Jianxin.Shi@nih.gov

Abstract

In prospective genomic studies (e.g., DNA methylation, metagenomics, and transcriptomics), it is crucial to estimate the overall fraction of phenotypic variance (OFPV) attributed to the high-dimensional genomic variables, a concept similar to heritability analyses in genome-wide association studies (GWAS). Unlike genetic variants in GWAS, these genomic variables are typically measured with error due to technical limitation and temporal instability. While the existing methods developed for GWAS can be used, ignoring measurement error may severely underestimate OFPV and mislead the design of future studies. Assuming that measurement error variances are distributed similarly between causal and noncausal variables, we show that the asymptotic attenuation factor equals to the average intraclass correlation coefficients of all genomic variables, which can be estimated based on a pilot study with repeated measurements. We illustrate the method by estimating the contribution of microbiome taxa to body mass index and multiple allergy traits in the American Gut Project. Finally, we show that measurement error does not cause meaningful bias when estimating the correlation of effect sizes for two traits. Overall, our study highlights the importance of accounting for measurement error in prospective genomic studies and provides a practical method for mitigating its impact on estimating OFPV.

Adaptive Aggregation for Longitudinal Quantile Regression Based on Censored History Process

Dianliang Deng

Department of Mathematics & Statistics, University of Regina, Canada

DianLiang.Deng@uregina.ca

Abstract

Most of studies for longitudinal quantile regression are based on the correct specification. Nevertheless, one specific model can hardly perform precisely under different conditions, and assessing which conditions are (approximately) satisfied to determine the optimal one is rather difficult. In the case of mixed effect model, the misspecification of the fixed effect part will cause a lack of predicting accuracy of random effects, and affect the efficiency of the cumulative function estimator. On the other hand, limit research has focused on incorporating multiple candidate procedures in longitudinal data analysis, which is of current emergency. This paper proposes an exponential aggregation weighting algorithm for longitudinal quantile regression. Based on the secondary smoothing loss function, we establish oracle inequalities for aggregated estimator. The proposed method is applied to evaluate the cumulative τ -th quantile function for additive mixed effect model with right-censored history process, an aggregation-based best linear prediction for random effects is constructed as well. We show that the asymptotic properties are conveniently imposed owing to the smoothing scheme. Simulation studies are carried out to exhibit the rationality, and our method is illustrated to analyze the data set from a multicenter automatic defibrillator implantation trial.

Developing and Validating Absolute Risk Prediction Models for Breast Cancer in Taiwan with Synthesized Data from Multiple Sources

Li-Hsin Chien[1]

[1]*Department of Applied Mathematics, Chung-Yuan Christian University*

lhchien@cycu.edu.tw

Abstract

Breast cancer incidence rates in Taiwan have been increasing rapidly in the past three decades, and it is the most common cancer among Taiwanese women. Based on some exploratory studies, the Taiwan Health Promotion Administration has implemented a universal biennial mammography screening (UBMS) program since 2004. However, it is essential to use a validated absolute risk model for breast cancer for individualized risk assessment regarding mammography screening. The purpose of the study is to develop and validate the breast cancer risk prediction models for the Taiwanese population. Based on the linkage of datasets from the UBMS from 2004 to 2019, Taiwan Cancer Registry (TCR) from 1979 to 2019, Taiwan Cause of Death Database (TCOD) from 1985 to 2019, and Taiwan National Health Insurance Research Database from 2000 to 2019, we developed and validated absolute risk prediction models for breast cancer among Taiwanese women aged 50—69. In fact, the linked dataset had a total of 1,746,580 women and was randomly divided into three disjoint datasets: one-half as the training set, one-quarter as the validation set, and the remaining quarter as the test set. Eventually, we obtained two models: one included mammography density, called the Taiwan Breast Cancer Model with mammography density (TBCM-M), and the other didn't, called the Taiwan Breast Cancer Model (TBCM). The other risk factors used included age at screening, age at menarche, age at menopause, parity, age at first birth, height, interaction between BMI and hormone replacement therapy (HRT) use, education, breast cancer family history in first-degree relatives, personal history of cancer, and breastfeeding. As a result, both models were well-calibrated, and TBCM-M (TBCM) had an AUC of 0.60 (0.59) for predicting breast cancer occurrence in the upcoming 5 years. In conclusion, both models could be used to improve the early detection of breast cancer. TBCM is applicable to women without any mammography screening record, and TBCM-M is suitable for women had mammography screening results. (Based on the work jointly done with Tzu-Yu Chen, Fang-Yu Tsai, I-Shou Chang, and Chao A. Hsiung.)

Keywords: Absolute risk prediction model; Synthesized data; Breast cancer.

Session I-4: Machine Learning
July 9, 2024 at 10:50-12:20
Yi-Xian Building 101 Conference Hall (050101)

◆ **Chair:**

I-Chen Lee (李宜真), Department of Statistics, National Cheng Kung University

◆ **Speakers:**

1. W.Y. Wendy Lou, Dalla Lana School of Public Health, University of Toronto, Canada
2. I-Ming Chiu (邱翊銘), Department of Economics, Rutgers University-Camden, U.S.A.
3. Chih-Hao Chang (張志浩), Department of Statistics, National Chengchi University

Fusion Clustering for Multi-Source Longitudinal Data

W.Y. Wendy Lou

Dalla Lana School of Public Health, University of Toronto, Canada

wendy.lou@utoronto.ca

Abstract

Motivated by a large-scale birth cohort with heterogeneous subject characteristics, we present a robust approach utilizing information from multiple sources over time to identify underlying phenotypes of the study population and their associated health outcomes. The pros and cons of the proposed approach will be discussed, and comparisons with some existing methods will be provided via two applications. Remaining challenges and possible strategies for analyzing similar types of data will also be presented.

Predicting Adolescent Depression in the U.S. Using a Two-Layered Ensemble Machine Learning Approach

I-Ming Chiu

Department of Economics, Rutgers University-Camden, U.S.A.

ichiu@camden.rutgers.edu

Abstract

According to Mental Health America (MHA), in 2022, "15.08% of youths aged 12-17 reported experiencing at least one major depressive episode (MDE). Furthermore, 10.6% of youths, equivalent to over 2.5 million individuals, are grappling with severe major depression." This poses a significant health concern because untreated adolescent mental health conditions may persist into adulthood, impacting both physical and mental well-being and limiting opportunities for leading fulfilling lives as adults. From an economic standpoint, MDE not only diminishes an individual's overall well-being but also diminishes overall productivity within a country.

In this study, using the cross-sectional National Survey on Drug Use and Health (NSDUH) data from 2011 to 2019, our primary objective was to investigate the association between severe depression in adolescents and three potential groups of contributing factors: sociodemographic characteristics (such as age, ethnicity, income, and family structure), parenting style, and school experiences, utilizing the logistic regression method. Our findings revealed that: Female adolescents face a higher risk than their male counterparts. The older age group (16-17) exhibits a greater risk compared to the younger age groups (14-15 and 12-13), respectively. Being Black or Asian/NHIPs is associated with a lower risk than being White. Those with Medicaid/CHIP, other insurance, or no insurance face a higher risk than those with private insurance. The absence of a father in the house, low parental involvement, and negative school experiences are all positively associated with a higher risk of depression in adolescents.

Furthermore, the logistic regression model served as a classifier to predict and identify adolescents with severe depression issues. Due to the imbalanced nature of the data (10% of depression cases), the respective accuracy and recall rates in the training data were 68.57% and 70.12% when the threshold value was set at 0.1. To enhance prediction performance, we introduced a second layer of classification using the decision trees algorithm. With this additional layer, we achieved an accuracy rate of 99.49% and a recall rate of 94.95%, a significant improvement on prediction. To mitigate overfitting, we evaluated the model's performance on the test data, where both accuracy and recall rates remained consistent. This predictive model facilitates early intervention and treatment to mitigate the consequences of depression among high-risk adolescent groups, consequently reducing both the social and economic costs associated with severe depression in adolescents.

Key Words: severe depression in adolescent; National Survey on Drug Use and Health; logistic classifier; decision trees, predictive model; confusion matrix; recall rate; accuracy rate.

Application of Machine Learning Techniques in Threshold Regression Modeling

Chih-Hao Chang

Department of Statistics, National Chengchi University

jhow@nccu.edu.tw

Abstract

This study introduces the threshold boundary regression (TBR) model for the analysis of datasets with binary or continuous responses. By integrating regression models with threshold boundary functions using explanatory variables, the TBR model constructs linear or nonlinear classifiers partitioning the responses into two groups, with separate regression models fitted to each group. To estimate the TBR model, we propose an ordered iterative algorithm called the TBR-WSVM algorithm. This algorithm combines weighted support vector machine (WSVM) techniques with maximum likelihood and least-squares methods. Through simulation studies and empirical analyses, we assess the performance of the TBR-WSVM algorithm. Our results indicate that the TBR-WSVM algorithm offers robust estimation and prediction capabilities for both linear and nonlinear threshold boundary models.

Keywords：Binary classification, Logistic regression, Nonlinear separability, Support vector machine, Threshold model

Parallel Session II

Session II-1: Biomedical Study
July 9, 2024 at 13:30-15:00
E. Sun Hall

◆ **Chair:**

Hong-Dar Isaac Wu (吳宏達), Department of Applied Mathematics, National Chung Hsing University

◆ **Speakers:**

1. Zhezhen Jin (金哲振), Department of Biostatistics, Columbia University, U.S.A.
2. Hua Zhou (周華), Department of Biostatistics, University of California, Los Angeles, U.S.A.
3. Jia-Han Shih (施嘉翰), Department of Applied Mathematics, National Sun Yat-sen University

Semiparametric Statistical Methods for the Analysis of Biomedical Data

Zhezhen Jin

Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, U.S.A.

zj7@cumc.columbia.edu

Abstract

In this talk, I will present statistical issues and challenges that I have encountered in my biomedical collaborative studies. It is very important to incorporate basic statistical principles and ideas in data analysis. The semiparametric statistical methods are robust and effective. It is essential to compare and identify biomarkers that are more informative to disease diagnosis and monitoring, and to evaluate various treatment procedure and plan on health outcome. After a discussion on the issues and challenges with real examples, I will review available statistical methods and present our newly developed semiparametric statistical methods that are useful for item reduction, differentiation of significant exposure factors and high dimensional data analysis.

Scalable and Robust Censored Linear Regression with Applications to Biobank Studies

Hua Zhou

Department of Biostatistics, University of California, Los Angeles (UCLA), U.S.A.

huazhou@ucla.edu

Abstract

We improve the synthetic variable approach for censored linear regression in three aspects. First, we introduce weighting to the estimation equation to improve estimation efficiency. Second, we extend the synthetic variable beyond right-censoring to the left- and interval-censored data. Third we derive synthetic variable for higher moments, extending its applicability to more complex models such as heterogeneous variances. We apply these scalable and robust censored linear regressions to the genetic studies of survival traits at biobank scale.

A Class of Regression Association Measure Based on Concordance

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

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

jhshih@math.nsysu.edu.tw

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

yhchen@stat.sinica.edu.tw

Abstract

Many regression association measures aiming at predictability of a dependent variable Y from an independent variable X have been studied recently. However, there is a lack of systematic discussion about these measures, including their rationale, properties, and estimation. In this talk, we introduce a class of measures which views the regression association of Y from X as the association between two independent replications from the conditional distribution of Y given X. The measures share a common form of the proportion of the variance of some function of Y that can be explained by X, rendering the measures a direct interpretation in terms of predictability. Moreover, the notion of two independent replications from the conditional distribution leads to a simple nonparametric estimation method based on the induced order statistics, hence no smoothing techniques are required. A gene data example is presented for illustration, showing that the considered measures can capture genes whose transcript levels exhibit oscillatory patterns in time.

Keywords: Concomitant, Functional association, Rank correlation

Session II-2: Survival Analysis
July 9, 2024 at 13:30-15:00
Room 210, the 2nd floor of College of Commerce (260210)

◆ **Chair:**

Yi-Ting Hwang (黃怡婷), Department of Statistics, National Taipei University

◆ **Speakers:**

1. Mei-Ling Ting Lee (丁美齡), Department of Epidemiology and Biostatistics, University of Maryland, College Park, U.S.A.
2. Jin Zhou (周瑾), Department of Biostatistics, University of California, Los Angeles, U.S.A.
3. Chung Chang (張中), Department of Applied Mathematics, National Sun Yat-sen University

Neural Network Extension of Threshold Regression for Event-Time Data

Mei-Ling Ting Lee

Department of Epidemiology and Biostatistics, University of Maryland, College Park, U.S.A.

mltlee@umd.edu

Abstract

Having the good property of estimators' collapsibility, the first-hitting-time based threshold regression with linear predictors can be easily used in causal inferences for time-to-event survival analysis. To extend the model to nonlinear cases, we consider a neural network extension of threshold regression which can efficiently model complex relationships among predictors and underlying health processes while providing clinically meaningful interpretations and tackle the challenge of capturing granular structure for high-dimensional data.

Scalable and Robust Joint Models for Longitudinal and Survival Outcomes

Jin Zhou

Department of Biostatistics, University of California, Los Angeles, U.S.A.

jinzhou@mednet.ucla.edu

Abstract

Healthcare data in the modern era offer a wealth of multi-level and multi-scale information over an extended period. These datasets present a unique chance to analyze disease progression and related time-varying risk factors, but existing statistical tools and algorithms for effectively analyzing exposure trajectories and disease onset at this scale are limited. In particular, the study of biomarker trajectories and their role in disease onset and progression is underdeveloped. In the first part of the talk, I will introduce TrajGWAS, a linear mixed model-based method for testing the genetic impact on a biomarker trajectory, including shifts in mean or within-subject variability. This method can handle biobank data with 100K to 1 million individuals and is robust against distributional assumptions. In the second part, I will present our recent efforts in developing a joint model for longitudinal and survival data that can handle biobank data with millions of subjects, intensive longitudinal measurements, and multiple random effects. Finally, I will showcase the application of these methods using Veterans Health Administration EHRs, covering 3.8 million veterans.

Heavy-Tailed Distribution for Combining Dependent P-Values with Asymptotic Robustness

Yusi Fang[1], Chung Chang*[2], Yongseok Park[1], George Tseng[3]

(* corresponding authors)

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

yuf31@pitt.edu

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

cchang@math.nsysu.edu.tw

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

ctseng@pitt.edu

Abstract

The issue of combining individual p-values to aggregate multiple small effects is a longstanding statistical topic. Many classical methods are designed for combining independent and frequent signals using the sum of transformed p-values with the transformation of light-tailed distributions, in which Fisher's method and Stouffer's method are the most well-known. In recent years, advances in big data promoted methods to aggregate correlated, sparse and weak signals; among them, Cauchy and harmonic mean combination tests were proposed to robustly combine p-values under unspecified dependency structure. Both of the proposed tests are the transformation of heavy-tailed distributions for improved power with the sparse signal. Motivated by this observation, we investigate the transformation of regularly varying distributions, which is a rich family of heavy-tailed distribution, to explore the conditions for a method to possess robustness to dependency and optimality of power for sparse signals. We show that only an equivalent class of Cauchy and harmonic mean tests has sufficient robustness to dependency in a practical sense. Moreover, a practical guideline to adjust significance level under dependency is provided based on our theorem and simulation. We also show an issue caused by large negative penalty in the Cauchy method and propose a simple, yet practical modification with fast computation. Finally, we present simulations and apply to a neuroticism GWAS application to verify the discovered theoretical insights.

Keywords: p-value combination

Session II-3: Time Series Analysis
July 9, 2024 at 13:30-15:00
Room 202, the 2nd floor of College of Commerce (260202)

◆ **Chair:**

Ie-Bin Lian (連怡斌), Department of Mathematics, National Changhua University of Education

◆ **Speakers:**

1. William W.S. Wei (魏武雄), Statistics, Operations, and Data Science, Temple University, U.S.A.
2. Li-Hsien Sun (孫立憲), Department of Statistics, National Central University
3. Hsin-Chieh Wong (翁新傑), Department of Statistics, National Taipei University

Issues on Time Series Analysis and Applications

William W.S. Wei

Department of Statistics, Operations and Data Science, Temple University, U.S.A.

wwei@temple.edu

Abstract

Time series are used in many studies and applications. In this presentation, we will begin with a univariate time series, consider aggregation and systematic sampling effects on time series analysis, and issues related to the use of time series and models. Since several time series are also often used in a study of the relationship of variables, we will also consider vector time series modeling and analysis. Although the basic procedures of model building between univariate time series and vector time series are the same, there are some important phenomena which are unique to vector time series. Moreover, time series can be seasonal and related to both space and time, so we will also discuss some issues related to multivariate seasonal vector time series and space time autoregressive moving average STARMA(p, q) models. Understanding these issues is important when we use time series data in modeling, analysis, and applications, regardless of whether it is a univariate or multivariate time series.

Keywords: time series, models, vector time series, aggregation

Interval-Based Time Series Analysis: Detecting Structural Shifts and Change-Point Estimation

Li-Hsien Sun

Department of Statistics, National Central University

tpsun7246@gmail.com

Abstract

We present a novel method for detecting structural shifts within interval-based time series data and accurately estimating change-points. To address this, we propose an innovative interval-based financial time series model that incorporates daily maximum, minimum, and terminal values, leveraging the geometric Brownian motion model. We then apply the proposed method to the financial time series where the data is emphasized by the significance of intra-daily information, including maximum and minimum prices while traditional finance models primarily focus on the daily closing price given the open price. We derive the likelihood function and corresponding maximum likelihood estimates (MLEs) using the Girsanov theorem and the Newton-Raphson (NR) algorithm. Through extensive simulations, we thoroughly evaluate the effectiveness of our proposed approach. Furthermore, empirical studies using real stock return data (S&P 500 index) from two critical periods, the 2008 financial crisis and the COVID-19 pandemic in 2020, allow us to assess its performance robustly.

Valid Post-Averaging Inference in AR-G/GARCH Models

Hsin-Chieh Wong

Department of Statistics & Fintech and Green Finance Center (FGFC), National Taipei University

hcwong@gm.ntpu.edu.tw

Abstract

Data analysis derives statistical inference from the resulting model performed by data-driven model (variable) selection or averaging. However, a puzzle is that inference after model selection may not guarantee to enjoy tests and confidence intervals provided by classical statistical theory. This paper proposes a valid post-averaging confidence interval in an AR model driven by a general GARCH model. To reach this goal, we investigate the asymptotic inference of the nested least squares averaging estimator under model uncertainty with fixed coefficients setup. Interestingly, based on a Mallows-type model averaging (MTMA) criterion, the weights of the under-fitted model decay to zero while only assigning the asymptotically random weights to the just-fitted and over-fitted models. Building on the asymptotic behavior of model weights, we derive the asymptotic distributions of the MTMA estimator. Monte Carlo simulations show that the proposed method achieves the nominal level.

Keywords: AR-G/GARCH, model averaging, heavy tails, tail behavior, stable distribution, Mallows-type criteria.

Session II-4: Biostatistics
July 9, 2024 at 13:30-15:00
Yi-Xian Building 101 Conference Hall (050101)

◆ **Chair:**

Chia-Hui Huang (黃佳慧), Department of Statistics, National Chengchi University

◆ **Speakers:**

1. Naisyin Wang (王乃昕), Department of Statistics, University of Michigan, U.S.A.
2. Li-Shan Huang (黃禮珊), Institute of Statistics, National Tsing Hua University
3. Dongdong Li (李東東), Department of Population Medicine, Harvard Medical School, U.S.A.

Utilizing Synthetic Components to Balance Privacy Protection and Data Utility

Naisyin Wang

Department of Statistics, University of Michigan, U.S.A.

nwangaa@umich.edu

Abstract

The importance of privacy protection is rising in the current practice of publicly sharing data. Different evaluation criteria in terms of privacy protection and data utilities are considered. They may or may not agree with each other. In this presentation, we illustrate ways to utilize synthetic components to balance privacy protection and data utilities for different evaluation criteria. One additional aim is to enable users easily implement statistical analysis using publicly shared data after the observations are processed with such privacy protection procedures. The efficacy and quality of the proposed procedures are illustrated theoretically and numerically via applications to biomedical datasets.

Promotion Time Cure Model with Local Polynomial Estimation

Li-Hsiang Lin[1], **Li-Shan Huang**[2]

[1]*Department of Mathematics and Statistics, Georgia State University,*

lhlin@gsu.edu

[2]*Institute of Statistics, National Tsing Hua University,*

lhuang@stat.nthu.edu.tw

Abstract

In modeling survival data with a cure fraction, flexible modeling of covariate effects on the probability of cure has important medical implications, which aids investigators in identifying better treatments to cure. This paper studies a semiparametric form of the Yakovlev promotion time cure model that allows for nonlinear effects of a continuous covariate. We adopt the local polynomial approach and use the local likelihood criterion to derive nonlinear estimates of covariate effects on cure rates, assuming that the baseline distribution function follows a parametric form. An algorithm is proposed to implement estimation at both the local and global scales. Asymptotic properties of local polynomial estimates, the nonparametric part, are investigated in the presence of both censored and cured data, and the parametric part is shown to be root-n consistent. The proposed methods are illustrated by simulated and real data with discussions on the practical applications of the proposed method, including the selections of the bandwidths in the local polynomial approach and the parametric baseline distribution family. Extension of the proposed method to multiple covariates are also discussed.

Keywords: Local likelihood; Proportional hazards model; Survival analysis

Evaluating Association Between Two Event Times with Observations Subject to Informative Censoring

Dongdong Li

Department of Population Medicine, Harvard Medical School, U.S.A.

dongdong_li@hphci.harvard.edu

Abstract

This work is concerned with evaluating the association between two event times without specifying the joint distribution parametrically. This is particularly challenging when the observations on the event times are subject to informative censoring due to a terminating event such as death. There are few methods suitable for assessing covariate effects on association in this context. We link the joint distribution of the two event times and the informative censoring time using a nested copula function. We use flexible functional forms to specify the covariate effects on both the marginal and joint distributions. In a semiparametric model for the bivariate event time, we estimate simultaneously the association parameters, the marginal survival functions, and the covariate effects. A byproduct of the approach is a consistent estimator for the induced marginal survival function of each event time conditional on the covariates. We develop an easy-to-implement pseudolikelihood-based inference procedure, derive the asymptotic properties of the estimators, and conduct simulation studies to examine the finite-sample performance of the proposed approach. For illustration, we apply our method to analyze data from the breast cancer survivorship study that motivated this research.

Parallel Session III

Session III-1: Industrial Statistics
July 9, 2024 at 15:20-16:50
E. Sun Hall

◆ **Chair:**

Sheng-Tsaing Tseng (曾勝滄), Institute of Statistics, National Tsing Hua University

◆ **Speakers:**

1. Dennis KJ Lin (林共進), Department of Statistics, Purdue University, U.S.A.
2. Tsai-Hung Fan (樊采虹), Graduate Institue of Statistics, National Central University
3. I-Tang Yu (俞一唐), Department of Statistics, TungHai University

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

Dennis K.J. Lin

Department of Statistics, Purdue University, West Lafayette, IN, U.S.A.

dkjlin@purdue.edu

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 is to find out "how can humans use AI, BI, and SI together to do things better."

General Random-Effects Trend Renewal Processes with Applications

Tsai-Hung Fan and Po-Shan Tseng

Graduate Institute of Statistic, National Central University

thfanncu@gmail.com

Abstract

A repairable system can be reused after repairs, but data from such systems often exhibit cyclic patterns. However, as seen in the charge-discharge cycles of a battery where capacity decreases with each cycle, the system's performance may not fully recover after each repair. To address this issue, the trend renewal process (TRP) transforms periodic data using a trend function to ensure the transformed data displays independent and stationary increments. This study investigates random-effects models with a conjugate structure, achieved by reparameterizing the TRP models. These random-effects TRP models, adaptable to any TRP model with a renewal distribution possessing a conjugate structure, provide enhanced convenience and flexibility in describing sample heterogeneity. Moreover, in addition to analyzing aircraft cooling system data, the proposed random-effects models are extended to accelerated TRP for assessing the reliability of lithium-ion battery data.

Keywords: Repairable system, accelerated trend renewal process, inverse Gaussian distribution, link function, end of performance (EOP).

Increment Degradation Model: A Bayesian Perspective

I-Tang Yu

Department of Statistics, Tunghai University

ityu@thu.edu.tw

Abstract

One frequently employed approach for describing the degradation phenomenon involves the use of a degradation model that relies on stochastic processes. In a stochastic-process-based degradation model, it is assumed that the increments follow a distribution with the additivity property. This property makes the further inferences mathematically and statistically tractable. However, it limits the choices of the distributions. This work aims to use those distributions without the additivity property to model the increments. Under the frame of Bayesian analysis, Markov Chain Monte Carlo algorithms are developed for executing the necessary computations. Given that the proposed degradation models do not adhere to the additivity property, this paper tackles the challenges involved in predicting the lifetime of both on-line and off-line products. The suitability of the proposed model is finally validated through a simulation study.

Session III-2: Causal Inference and Lifetime Data Analysis
July 9, 2024 at 15:20-16:50
Room 210, the 2nd floor of College of Commerce (260210)

◆ **Chair:**

Tsung-Shan Tsou (鄒宗山), Graduate Institute of Statistics, National Central University

◆ **Speakers:**

1. Jialiang Li (栗家量), Department of Statistics & Data Science, National University of Singapore
2. Yong Chen (陳勇), Department of Biostatistics, University of Pennsylvania, U.S.A.
3. Hsin-wen Chang (張馨文), Institute of Statistical Sciences, Academia Sinica

Efficient Auxiliary Information Synthesis for Cure Rate Model

Jialiang Li

Department of Statistics & Data Science, National University of Singapore

jialiang@nus.edu.sg

Abstract

We propose a new auxiliary information synthesis method to utilize subgroup survival information at multiple time points under the semiparametric mixture cure rate model. After summarizing the auxiliary information via estimating equations, a control variate technique is adopted to reduce the variance efficiently, together with a test statistic to check the homogeneity assumption. Revision using penalization is further considered to adaptively accommodate potential population heterogeneity. Our methods can be adjusted when the uncertainty is not negligible. We establish asymptotic properties of our proposed estimators, and demonstrate their practical performances through extensive simulations and an invasive breast cancer study.

Real-World Effectiveness of BNT162b2 Against Infection in Children: Causal Inference Under Misclassification in Treatment Status

Yong Chen

Department of Biostatistics, University of Pennsylvania, U.S.A.

ychen123@pennmedicine.upenn.edu

Abstract

The current understanding of the long-term effectiveness of the BNT162b2 vaccine for a range of outcomes across diverse U.S. pediatric populations is limited. In this study, we assessed the effectiveness of BNT162b2 against various strains of the SARS-CoV-2 virus using data from a national collaboration of pediatric health systems (PEDSnet). In the U.S., immunization records are often captured and stored across multiple disconnected sources, resulting in incomplete vaccination records in patients' electronic health records (EHR). We developed a novel trial emulation framework that accounts for misclassification bias in vaccine documentation in EHRs. The effectiveness of the BNT162b2 vaccine was estimated with respect to adolescents aged 12 to 20 years during the Delta variant period, children aged 5 to 11 years, and adolescents aged 12 to 20 years during the Omicron variant period.

Bivariate Analysis of Distribution Functions Under Biased Sampling

Hsin-wen Chang

Institute of Statistical Science, Academia Sinica

hwchang@stat.sinica.edu.tw

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 COVID-19 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.

Session III-3: Experimental Designs
July 9, 2024 at 15:20-16:50
Room 202, the 2nd floor of College of Commerce (260202)

◆ **Chair:**

Mong-Na Lo Huang (羅夢娜), Department of Applied Mathematics, National Sun Yat-sen University

◆ **Speakers:**

1. Ming-Hung Kao (高銘宏), School of Mathematical and Statistical Sciences, Arizona State University, U.S.A.
2. Hsiang-Ling Hsu (許湘伶), Institute of Statistics, National University of Kaohsiung
3. Cheng-Yu Sun (孫誠佑), Institute of Statistics, National Tsing Hua University

Optimal Study Designs for Sparse Functional Data Analysis

Ming-Hung (Jason) Kao

School of Mathematical and Statistical Sciences, Arizona State University, U.S.A.

ming-hung.kao@asu.edu

Abstract

Functional data analysis (FDA) is powerful in extracting useful information from underlying random functions of interest. Similar to other data analysis methods, having high-quality data is key to the success of statistical inference with FDA, and the importance of judiciously selecting a good study design to collect informative data cannot be overemphasized. Here, we propose optimal study design methods for FDA with a focus on sparse functional data, where noisy observations from random curves are collected at sparse, possibly irregularly spaced, time points. Theoretical results and computational methods are developed, and their usefulness on obtaining optimal study designs for sparse FDA is demonstrated via some simulation studies and real examples.

Optimal Designs with Multiple Correlated Responses for Experiments with Mixtures

Hsiang-Ling Hsu

Institute of Statistics, National University of Kaohsiung

hsuhl@nuk.edu.tw

Abstract

A mixture experiment within the (q-1)-dimensional probability simplex is a specific experimental setup in which the q factors are non-negative and adhere to the sum of all factors equals one. In this talk, we investigate the issue of the optimal approximate designs with the k-correlated response mixture experimental models. In the multiple correlated response mixture models, we explore the improvement design class, known as the complete class, in relation to the Kiefer ordering for a given design. Based on the complete class results, we delve into the properties of optimal designs for multiresponse models using the well-established equivalence theorem. For specific multiresponse model settings under the D-optimal design criterion, the optimal results can reduce multiresponse experimental design problems to single-response experimental design problems. An illustrative example showcasing optimal designs for two correlated response mixture experimental models is presented.

Keywords: Design optimality; Invariant design; Kiefer ordering; Mixture experimental model; Weighted centroid design.

Space-Filling Regular Designs Under a Minimum Aberration-Type Criterion

Cheng-Yu Sun

Institute of Statistics, National Tsing Hua University

chengyus@stat.nthu.edu.tw

Abstract

Space-filling designs plays a vital role in computer experiments. Common criteria for selecting such designs are either distance- or discrepancy-based. Recently, Tian and Xu introduced a minimum aberration-type criterion known as the Space-Filling Pattern (SFP). This criterion examines whether a design exhibits stratifications on a series of grids, and can effectively distinguish strong orthogonal arrays of same strengths. Subsequently, Shi and Xu refined the SFP to the stratification pattern (SP). They showed that designs excelling under the SFP construct better surrogate models than those meeting many other uniformity criteria. In this study, we provide a new justification for both SFP and SP, and discuss a new pattern that is similar to the SP. Then, our focus shits to the construction of space-filling regular designs. We show that both the SP and our proposed pattern of a regular design can be determined by counting different types of words of given lengths. This result allows for a complete search for the most space-filling regular designs of moderate run sizes.

Session III-4: Deep Learning and High Dimensional Data Analysis
July 9, 2024 at 15:20-16:50
Yi-Xian Building 101 Conference Hall (050101)

◆ **Chair:**

Yuan, Juan-Ming (袁淵明), Department of Data Science and Big Data Analytics, Providence University

◆ **Speakers:**

1. Tai-Been Chen (陳泰賓), Department of Radiological Technology, Faculty of Medical Technology, Teikyo University, Tokyo, Japan
2. Yen-Lung Tsai (蔡炎龍), Office of Student Affairs, National Chengchi University
3. An-Shun Tai (戴安順), Department of Statistics, National Cheng Kung University

Deep Learning Applications in Chest X-Ray Classification, CT Liver Tumors Segmentation, and Detection of Stenosis on X-Ray Coronary Angiography

Tai-Been Chen

Department of Radiological Technology, Faculty of Medical Technology, Teikyo University, Tokyo, Japan

ctb@isu.edu.tw

Abstract

This speech explores the transformative impact of deep learning in medical imaging, focusing on three pivotal applications: chest X-ray classification, computed tomography (CT) liver tumors segmentation, and the detection of stenosis in X-ray coronary angiography.

In the realm of chest X-ray classification, we utilized a hybrid artificial intelligence model, Fusion Convolutional Neural Network (CNN), which integrates five distinct CNN architectures following transfer learning. This model was trained on a dataset of 5,260 images, comprising 1,792 normal, 1,658 COVID-19, and 1,800 bacterial pneumonia images. The Fusion CNN model demonstrated exemplary performance with an accuracy of 99.4% and a Kappa value of 99.1%.

For CT liver tumor segmentation, we employed fully convolutional networks (FCN) with backbones such as Xception, InceptionResNetv2, MobileNetv2, ResNet18, and ResNet50. Analyzing 7,190 2D CT images from 131 patients, the best results were achieved using ResNet50 in FCN, with metrics such as global accuracy at 99.9%, mean intersection over union (IoU) at 95.4%, and Weighted IoU at 99.8%.

Lastly, the detection of stenosis on X-ray coronary angiography was performed using a real-time YOLO model on video-derived static images. From 120 patients, 2,708 images were processed, achieving an optimal IoU of 78.8% with the detection speed approximating 24 frames per second using ResNet-50.

These case studies highlight the efficacy of advanced deep learning models in enhancing diagnostic accuracies and streamlining medical imaging workflows, presenting a significant advancement in healthcare technology.

Key words: Chest X-ray Classification, CT Liver Tumor Segmentation, X-ray Coronary Angiography, Convolutional Neural Networks (CNN), Fully Convolutional Networks (FCN), YOLO (You Only Look Once)

Contrastive Learning for Time Series Data: Predicting Stock Price Movements

Yen-Lung Tsai

Department of Mathematical Sciences, National Chengchi University

yenlung@nccu.edu.tw

Abstract

This talk explores the application of contrastive learning in deep learning to analyze time series data, with a specific focus on predicting stock price movements. We begin by introducing the fundamental concepts of deep learning, emphasizing the view that each hidden layer can be regarded as a form of feature engineering, and how feature representations can be learned through pretext tasks. Subsequently, we delve into the concepts of self-supervised learning and contrastive learning, elucidating the application of contrastive learning in time series data, particularly in the context of predicting stock price fluctuations. The talk will include our research applying contrastive learning on real stock market data. Lastly, we will discuss potential future research directions and the outlook for applications in the financial domain.

Robust and Flexible High-Dimensional Causal Mediation Model for DNA Methylation Studies

An-Shun Tai

Department of Statistics, National Cheng Kung University

ashtai@gs.ncku.edu.tw

Abstract

In the pathogenesis of diseases, DNA methylation (DNAm) markers play a pivotal role in influencing gene expression and engaging in diverse biological processes. Given the extensive number of DNAm markers, exceeding half a million, implementing a high-dimensional mediation model is necessary to identify the activated DNAm markers within the mediation pathway and assess their mediation effects. Most existing high-dimensional mediation models necessitate stringent assumptions, including correctly prespecifying the mediation relationship and determining all outcomes, mediators, and exposure models. However, fulfilling these assumptions is challenging in the context of high-dimensional mediators. This study introduces a novel Bayesian estimation procedure for interventional mediation effects, offering robustness against model misspecification and flexibility in prespecifying the mediation structure. Spike-and-slab priors are employed to integrate Bayesian variable selection into the modeling process. The proposed method is demonstrated using publicly available genome-wide array-based cancer studies to estimate the causal effects mediated through DNAm.

Parallel Session IV

Session IV-1: Functional Data Analysis & Dimension Reduction
July 10, 2024 at 10:30-12:00
E. Sun Hall

◆ **Chair:**

Mei-Hui Guo (郭美惠), Department of Applied Mathematics, National Sun Yat-sen University

◆ **Speakers:**

1. Hans-Georg Müller, Department of Statistics, University of California, Davis, U.S.A.
2. Ci-Ren Jiang (江其衽), Institute of Statistics and Data Science, National Taiwan University
3. Lih-Yuan Deng (鄧利源), Department of Mathematical Sciences, University of Memphis, U.S.A.

Quantifying Variation for Random Objects Via Distance Profiles

Hans-Georg Müller

Department of Statistics, University of California, Davis, U.S.A.

hgmueller@ucdavis.edu

Abstract

For random objects, i.e., data taking values in a general separable metric space (X,d) with a probability measure, the distance profile of an element x in the space refers to the distribution of the distances between x and the other elements of X. Since depth profiles are one-dimensional distributions, optimal transports between them are easily obtained and can be harnessed to define transport ranks, which capture the centrality of each element in X with respect to the entire data cloud. We study the properties of these transport ranks and show that they provide an effective device for detecting and visualizing patterns in samples of random objects and that empirical estimates converge to their population targets. We demonstrate the usefulness of transport ranks for data analysis tasks involving distributional data, compositional data and network data. This talk is based on joint work with Yaqing Chen (Rutgers) and Paromita Dubey (USC).

Eigen-Adjusted FPCA

Ci-Ren Jiang

Institute of Statistics and Data Science, National Taiwan University

cirenjiang@ntu.edu.tw

Abstract

Functional Principal Component Analysis (FPCA) has become a widely used dimension reduction tool for functional data analysis. When additional covariates are available, existing FPCA models integrate them either in the mean function or in both the mean function and the covariance function. However, methods of the first kind are not suitable for data that display second-order variation, while those of the second kind are time-consuming and make it difficult to perform subsequent statistical analyses on the dimension-reduced representations. To tackle these issues, we introduce an eigen-adjusted FPCA model that integrates covariates in the covariance function only through its eigenvalues. In particular, different structures on the covariate-specific eigenvalues—corresponding to different practical problems—are discussed to illustrate the model's flexibility as well as utility. To handle functional observations under different sampling schemes, we employ local linear smoothers to estimate the mean function and the pooled covariance function, and a weighted least square approach to estimate the covariate-specific eigenvalues. The convergence rates of the proposed estimators are further investigated under the different sampling schemes. In addition to simulation studies, the proposed model is applied to functional Magnetic Resonance Imaging scans, collected within the Human Connectome Project, for functional connectivity investigation. Supplementary materials for this article are available online.

Big Data Model Building Using Dimension Reduction and Sample Selection

Lih-Yuan Deng

Department of Mathematical Sciences, University of Memphis, U.S.A.

lihdeng@memphis.edu

Abstract

It is difficult to handle the extraordinary data volume generated in many fields with current computational resources and techniques. This is very challenging when applying conventional statistical methods to big data. A common approach is to partition full data into smaller subdata for purposes such as training, testing, and validation. The primary purpose of training data is to represent the full data. To achieve this goal, the selection of training subdata becomes pivotal in retaining essential characteristics of the full data. Recently, several procedures have been proposed to select "optimal design points" as training subdata under pre-specified models, such as linear regression and logistic regression. However, these subdata will not be "optimal" if the assumed model is not appropriate. Furthermore, such subdata cannot be useful to build alternative models because it is not an appropriate representative sample of the full data. In this paper, we propose a novel algorithm for better model building and prediction via a process of selecting a "good" training sample. The proposed subdata can retain most characteristics of the original big data. It is also more robust that one can fit various response model and select the optimal model.

Session IV-2: Recent Advances in Risk Analysis
July 10, 2024 at 10:30-12:00
Room 210, the 2nd floor of College of Commerce (260210)

◆ **Chair:**

Jun Zhao, International Chinese Statistical Association, U.S.A.

◆ **Speakers:**

1. Fabrizio Ruggeri, Institute of Applied Mathematics and Information Technology, Italian National Research Council, Italia
2. Kyoji Furukawa, Biostatistics Center, Kurume University, Japan
3. Rachel Huang (黃瑞卿), Department of Finance, National Central University

Recent Advances in Adversarial Risk Analysis

Fabrizio Ruggeri

CNR (Italian National Research Council) Senior Fellow, Italia

fabrizio@mi.imati.cnr.it

Abstract

In the talk I will present some of my recent works in the field of Adversarial Risk Analysis. I will talk about Adversarial Classification. In multiple domains such as malware detection, automated driving systems, or fraud detection, classification algorithms are susceptible to being attacked by malicious agents willing to perturb the value of instance covariates in search of certain goals. Such problems pertain to the field of adversarial machine learning and have been mainly dealt with, perhaps implicitly, through game-theoretic ideas with strong underlying common knowledge assumptions. These are not realistic in numerous application domains in relation to security. I will present an alternative statistical framework that accounts for the lack of knowledge about the attacker's behavior using adversarial risk analysis concepts.

Statistical Challenges in Radiation Risk Assessment

Kyoji Furukawa

Biostatistics Center, Kurume University, Japan

furukawa_kyoji@kurume-u.ac.jp

Abstract

While radiation is essential and beneficial in medicine for diagnostic and therapeutic purposes, the potential risk of radiation to human health has been of great public concerns. The risks derived from the epidemiological cohort study of Japanese atomic-bomb survivors (Life Span Study; LSS) have been considered as the most reliable sources of information to understand the nature of radiation effects on human health. They have been frequently utilized for various purposes, e.g., to recommend radiation protection standards for the general public, to project long-term health effects in exposed populations, and to evaluate the benefits of medical exposure relative to potential health risks.

Generally, risk assessments based on survival data observed in a large-scale cohort study can suffer from a number of complex statistical issues. Data often involves various types of incompleteness, such as missing data, measurement error or misclassification, and confounding, each of which can introduce bias and loss of efficiency in estimation of the exposure effect. In a lifetime-long follow-up, censoring due to competing risk events can be often non-independent, and unobservable heterogeneity can cause unexpected impacts thorough selection of subjects at less and less risks overtime. Despite accumulation of data over the decades, uncertainties in models for radiation-associated risks (dose response shape, effect modification) are still not small. Each of these issues is suspected to be a part of the reasons why understanding remains limited about the risk associated with exposure at low doses and low dose rates, which are most relevant to exposures of our concerns today.

In this talk, focusing on studies of the atomic-bomb survivors, recent developments of statistical approaches to various issues in radiation risk assessment are introduced. Furthermore, future research directions needed to improve statistical methodologies to perform defendable risk assessment and help understand the mechanism of radiation-induced adverse health effects are also discussed.

Keywords: Radiation epidemiology, risk assessment, survival analysis, incomplete data

Hedge Funds Performance: Are Crypto Hedge Funds the Rising Star?

Rachel Huang

Department of Finance, National Central University

rachelhuang.ncu@gmail.com

Abstract

This study compares the performance of Crypto hedge funds to conventional strategies using a non-parametric and utility-based measure called almost stochastic dominance. The measure is a criterion for ranking distributions for most economically important investors. Two new performance indices consistent with this measure is. Analyzing data from July 2013 to July 2022, both the criterion and our indices reveal that Crypto hedge funds outperform others over one to three years. Furthermore, our indices suggest that Equity Hedge, Risk Parity, and Relative Value strategies outperform Event-Driven, Fund of Funds, and Macro strategies for investment horizons ranging from one to three years.

Keywords: Hedge Funds, Almost Stochastic Dominance, Performance Index, Cryptocurrency, Investment Strategy

Session IV-3: Cutting-Edge Statistical Modeling Approaches for Multifaceted Data
July 10, 2024 at 10:30-12:00
Room 202, the 2nd floor of College of Commerce (260202)

◆ **Chair:**

Tsung-I Lin (林宗儀), Institute of Statistics, National Chung Hsing University

◆ **Speakers:**

1. Victor Hugo Lachos, Department of Statistics, University of Connecticut, U.S.A.
2. Mohammad Arashi, Department of Mathematical Sciences, Ferdowsi University of Mashhad, IRAN
3. Chang-Yun Lin (林長鋆), Institute of Statistics, National Chung Hsing University

Heckman Selection Contaminated Normal Model: Parameter Estimation via the EM-Algorithm

Victor H. Lachos

Department of Statistics, University of Connecticut, Storrs, U.S.A.

hlachos@uconn.edu

Abstract

The Heckman selection model is perhaps the most popular econometric model in the analysis of data with sample selection. The analyses of this model are based on the normality assumption for the error terms, however, in some applications, the distribution of the error term departs significantly from normality, for instance, in the presence of heavy tails and/or atypical observation. In this paper, we propose a novel Heckman selection model where the random errors follow a bivariate contaminated normal distribution. We develop an analytically tractable and efficient EM-type algorithm for iteratively computing maximum likelihood estimates of the parameters, with standard errors as a by-product. The algorithm has closed-form expressions at the E-step, that rely on formulas for the mean and variance of the truncated contaminated normal distributions. Simulation studies show the vulnerability of the Heckman selection-normal model, as well as the robustness aspects of the Heckman selection-contaminated normal model. Two real examples are analyzed, illustrating the usefulness of the proposed methods. The proposed algorithms and methods are implemented in the new R package HeckmanEM.

Keywords: EM-type algorithms, Heckman selection model, Multivariate contaminated normal, Robustness.

High-dimensional Regression Analysis with Machine Learning

Mohammad Arashi

Department of Statistics, Ferdowsi University of Mashhad, IRAN

m_arashi_stat@yahoo.com

Abstract

In the present era, thanks to advancements in technology, accessing data with abundant features has become commonplace. As a result, feature engineering plays a crucial role in data analysis. When dealing with high-dimensional data problems, where the number of features (p) surpasses the number of samples (n), addressing multicollinearity poses a challenge in model estimation. This paper introduces a novel and uncomplicated method for estimating regression parameters in scenarios characterized by large dimensions and multicollinearity. The proposed method capitalizes on the advantageous characteristics of machine learning and the straightforward structure of a class of linear unified estimators. In situations where multicollinearity is prevalent and p exceeds n, our approach offers a simple and remarkably fast means of estimating regression coefficients. The superior performance of our proposed method is evident from numerical examinations.

Keywords: Chemometrics, High-dimensional regression, Machine learning, Multicollinearity, Ridge and Liu Regressions

Design Construction and Model Selection for Small Mixture-Process Variable (MPV) Experiments with High-Dimensional Model Terms

Kashinath Chatterjee and Chang-Yun Lin*

Department of Population Health Sciences, Division of Biostatistics and Data Sciences, Augusta University, Augusta, Georgia, U.S.A.

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

chlin6@nchu.edu.tw

Abstract

This paper considers the design construction and model selection for mixture-process variable experiments where the number of variables is large. For such experiments the generalized least squares estimates cannot be obtained and hence it will be difficult to identify the important model terms. To overcome these problems, here we employ the generalized Bayesian-D criterion to choose the optimal design and apply the Bayesian analysis method to select the best model. Two algorithms are developed to implement the proposed methods. A fish-patty experiment demonstrates how the Bayesian approach can be applied to a real experiment. Simulation studies show that the proposed method has a high power to identify important terms and well controls the type I error.

Keywords: Bayesian, D-optimality criterion, generalized Bayesian-D criterion, split-plot design.

Session IV-4: Experimental Designs
July 10, 2024 at 10:30-12:00
Yi-Xian Building 101 Conference Hall (050101)

◆ **Chair:**

Ray-Bing Chen (陳瑞彬), Department of Statistics, National Cheng Kung University

◆ **Speakers:**

1. Weng Kee Wong (王永琪), Department of Biostatistics, University of California, Los Angeles, U.S.A.
2. Qian Helen Li, StatsVita, LLC, U.S.A.
3. Ming-Chung Chang (張明中), Institute of Statistical Science, Academia Sinica

Optimal Exact Designs for Small Studies in Toxicology with Applications to Hormesis via Metaheuristics

Weng Kee Wong, Mr. Brain Wu, Ray-Bing Chen

Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, California

wkwong@ucla.edu

Department of Statistics, National Cheng Kung University

Department of Statistics, National Cheng Kung University

Abstract

There are theory-based methods for searching model-based optimal continuous designs when the sample size is large. The elegant theory can also confirm optimality of any design and provide an assessment of how efficient the design is without knowing the optimum. When the sample size is small, the theory may no longer apply, and even if it did, the optimal continuous design may not be implementable. We propose a nature-inspired metaheuristic algorithm to find efficient exact designs for experiments with a small sample size. As an application, we find efficient exact designs for detecting existence of hormesis in a hormetic model, and to estimate the hormesis threshold.

We demonstrate the flexibility of the algorithm by finding locally D-optimal exact designs for estimating model parameters in logistic models for small experiments and other types of optimal exact designs for longitudinal studies with a few time points.

Extending the Use of Adaptive Design Beyond Sample Size Re-Estimation

Qian H. Li

StatsVita, U.S.A.

qianhelenli@gmail.com

Abstract

Adaptive design for sample size re-estimation has become a popular design method in clinical trials to test innovative interventions. In studies testing treatments for rare diseases, the trials may face uncertainty in needed treatment duration to demonstrate treatment effect during long term follow-up. The time points at which the primary efficacy endpoint should be evaluated may need to be adapted based on the interim evaluation of the treatment effect. In this presentation, we present applications of extending the adaptive design from sample size adaptation, which is proposed by Mehta and Pocock (2011) to adaptations in efficacy evaluation timepoints by identifying promising zones for conditional power calculated at interim. We will demonstrate the adequate control of type I error in such applications. Simulation studies are used to illustrate the operating characteristics of the adaptive design and decision rules based on conditional power.

Orthogonalized Moment Aberration for Multi-Stratum Factorial Designs

Ming-Chung Chang

Institute of Statistical Science, Academia Sinica

mcchang@stat.sinica.edu.tw

Abstract

Multi-stratum factorial designs are prevalent in industrial and agricultural applications. However, the experimental units in these designs are subject to multiple error sources, posing a challenge in designing optimal multi-stratum factorial designs. This presentation introduces an orthogonalized moment aberration criterion for multi-stratum factorial designs. This row-based criterion enables fast computation. Additionally, we establish that the minimum orthogonalized moment aberration designs are optimal under the criterion proposed by Chang and Cheng [Ann. Stat. 46 (2018) 1779-1806].

Parallel Session V

Session V-1: Statistics and Data Science
July 10, 2024 at 13:00-14:30
E. Sun Hall

◆ **Chair:**

Chien-Tai Lin (林千代), Department of Mathematics, Tamkang University

◆ **Speakers:**

1. Samuel Kou (寇星昌), Department of Statistics, University of Harvard, U.S.A.
2. Jung-Ying Tzeng (曾仲瑩), Department of Statistics and Bioinformatics Research Center, North Carolina State University, U.S.A.
3. Hsiuying Wang (王秀瑛), Institute of Statistics, National Yang Ming Chiao Tung University

Statistical Inference of Dynamic Systems via Manifold-Constrained Gaussian Processes

Samuel Kou

Department of Statistics, Harvard University, U.S.A.

kou@stat.harvard.edu

Abstract

Parameter estimation for nonlinear dynamic system models, represented by ordinary differential equations (ODEs), using noisy and sparse data is a vital task in many fields. We will introduce a fast and accurate method, MAGI (Manifold-constrained Gaussian process Inference), in this task. MAGI uses a Gaussian process model over time-series data, explicitly conditioned on the manifold constraint that derivatives of the Gaussian process must satisfy the ODE system. By doing so, we completely bypass the need for numerical integration and achieve substantial savings in computational time. MAGI is also suitable for inference with unobserved system components, which often occur in real experiments. MAGI is distinct from existing approaches as we provide a principled statistical construction under a Bayesian framework, which incorporates the ODE system through the manifold constraint. We demonstrate the accuracy and speed of MAGI using realistic examples based on physical experiments.

Transfer Learning with False Negative Control Improves Polygenic Risk Prediction

Xinge Jessie Jeng[1], Yifei Hu[2], Vaishnavi Venkat[3], Tzu-Pin Lu[4], Jung-Ying Tzeng[5]

[1]*Department of Statistics, North Carolina State University, Raleigh, North Carolina, U.S.A.*

jessie_jeng@ncsu.edu

[2]*Department of Statistics, North Carolina State University, Raleigh, North Carolina, U.S.A.*

hyifei@ncsu.edu

[3]*Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, U.S.A.*

vvenka23@ncsu.edu

[4]*Institute of Health Data Analytics and Statistics, National Taiwan University, Taipei, Taiwan*

tplu@ntu.edu.tw

[5]*Department of Statistics and Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, U.S.A.*

jytzeng@ncsu.edu

Abstract

Polygenic risk score (PRS) is a quantity that aggregates the effects of variants across the genome and estimates an individual's genetic predisposition for a given trait. PRS analysis typically contains two input data sets: base data for effect size estimation and target data for individual-level prediction. Given the availability of large-scale base data, it becomes more common that the ancestral background of base and target data do not perfectly match. In this paper, we treat the GWAS summary information obtained in the base data as knowledge learned from a pre-trained model, and adopt a transfer learning framework to effectively leverage the knowledge learned from the base data that may or may not have similar ancestral background as the target samples to build prediction models for target individuals. Our proposed transfer learning framework consists of two main steps: (1) conducting false negative control (FNC) marginal screening to extract useful knowledge from the base data; and (2) performing joint model training to integrate the knowledge extracted from base data with the target training data for accurate trans-data prediction. This new approach can significantly enhance the computational and statistical efficiency of joint-model training, alleviate over-fitting, and facilitate more accurate trans-data prediction when heterogeneity level between target and base data sets is small or high.

Keywords: genetic risk score (GRS); cross-ethnic PRS; ancestry-diverse PRS.

Tolerance Interval for COVID-19 Data Prediction

Hsiuying Wang

Institute of Statistics, National Yang Ming Chiao Tung University

wang@stat.nycu.edu.tw

Abstract

Tolerance intervals (TIs) are widely used in various applications including manufacturing engineering, clinical research, and the pharmaceutical industry. TIs can be used to construct limits on control charts for monitoring quality characteristics or to establish upper bounds on quantities. COVID-19, which is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), emerged in Wuhan, China, in December 2019 and rapidly grew into a worldwide pandemic.

In this study, the TI method is applied to develop a control chart for monitoring the daily case count of COVID-19 and to set upper limits on case numbers. Real data examples demonstrate the superiority of these approaches, which can aid in pandemic management.

Session V-2: Deep Learning and AI
July 10, 2024 at 13:00-14:30
Room 210, the 2nd floor of College of Commerce (260210)

◆ **Chair:**

Chun-Shu Chen (陳春樹), Institute of Statistics, National Central University

◆ **Speakers:**

1. Fushing Hsieh (謝復興), Department of Statistics, University of California, Davis, U.S.A.
2. Ting-Li Chen (陳定立), Institute of Statistical Sciences, Academia Sinica
3. Guan-Hua Huang (黃冠華), Institute of Statistics, National Yang Ming Chiao Tung University

Data Intelligence vs A.I.

Fushing Hsieh

Department of Statistics, University of California, Davis, U.S.A.

fhsieh@ucdavis.edu

Abstract

Every data set derived from a complex system of interest is embedded with information content that encoded with data curator's and system's dynamic intelligence. In this talk, the Categorical Exploratory Data Analysis (CEDA) paradigm is applied to compute and extract such information content as fully as possible under constraints pertaining to dataset's sample size and dimensionality. This paradigm works for all datatypes and is free from all man-made assumptions and structures. We demonstrate how to synthesize bundles of computed pieces of information into visible and explainable pattern-based potential knowledge and intelligence. We then discuss the possibility of transforming and training language models (A.I.) that can coherently and effectively express such data intelligence by working with domain scientists. Finally, we discuss whether this computational approach is one efficient way of retaining and expanding experts' intelligence regarding any complex system of interest. Three real examples from Agronomy are discussed and illustrated.

Multiscale Major Factor Selections for Complex System Data with Structural Dependency and Heterogeneity

Ting-Li Chen*

Institute of Statistical Sciences, Academia Sinica

tlchen@stat.sinica.edu.tw

Abstract

This study delves into the intricate multiscale structures that underlie large complex systems, adopting a bottom-up approach to uncover the hidden heterogeneity and structural dependencies within structured data sets. We demonstrate our method using two real-world complex systems, showing how computed hierarchical structures, which show broken symmetry, can capture and represent the data's information content. This information is shown through graphs, providing an efficient way to tackle scientific challenges within the system that are usually hard to address. Our exploration and analytical processes are based on conditional entropy and mutual information assessments conducted on contingency table platforms, following the categorization of all quantitative features. We introduce Categorical Exploratory Data Analysis (CEDA) to initially identify global major factors that significantly interact with the targeted response variable in the context of structural dependency and multiple covariates. Each identified global major factor serves as a lens for heterogeneity, allowing us to partition the data set into sub-collections based on its categories. This 'deassociating' approach effectively reduces structural dependencies among features, enabling a refined selection of major factors at the sub-collection level that can reveal additional informative content beyond the global perspective. By synthesizing insights from multiple heterogeneity viewpoints, we present informative patterns that directly contribute to advancements in prediction, classification, and the detection of subtle dynamic shifts within the system.

Automated Convolutional Neural Network and Transformer for Multi-Class Classification of Three-Dimensional Brain Images

Cheng-Chun Wu and Guan-Hua Huang*

Institute of Statistics, National Yang Ming Chiao Tung University

ghuang@nycu.edu.tw

Abstract

In this study, we designed and developed multiple model architectures that can effectively be applied to three-dimensional (3D) single photon emission computed tomography (SPECT) images, addressing the multi-class classification task of predicting stages of Parkinson's disease. We treated the 3D SPECT imaging as a sequence of two-dimensional (2D) image slice sequences, and fed them into the 2D convolutional neural network (CNN) or Transformer model to extract features. Then, these features were summarized in the second layer that implemented attention-mechanism-based models (including Transformer) to take into account the absolute or relative position information of input the slice. In the end, summarized features were used to derive a forecast for the disease stage. Finally, we applied automated machine learning (AutoML) to pre-processing selection and hyperparameter optimization tasks, which can quickly achieve the goal through automation and reduce the labor involved in manual selection. The analysis results showed that using Transformer for both feature extraction and summarization did not meet expectations. However, combining CNN for feature extraction and Transformer for summarization significantly improved the prediction accuracy and F1 scores of 3D image classification. Additionally, AutoML identified a combination of pre-processing approaches and hyperparameters that yielded a deep learning model with performance comparable to the best manually selected model, saving time and effort.

Keywords: Parkinson's disease, single photon emission computed tomography, deep learning, Transformer, automated machine learning.

Session V-3: Statistics and Data Science
July 10, 2024 at 13:00-14:30
Room 202, the 2nd floor of College of Commerce (260202)

◆ **Chair:**

Henghsiu Tsai (蔡恆修), Institute of Statistical Sciences, Academia Sinica

◆ **Speakers:**

1. Ying Hung (洪瑛), Department of Statistics, Rutgers, the State University of New Jersey, U.S.A.
2. Takeshi Emura (江村剛志), Research Center for Medical and Health Data Science, the Institute of Statistical Mathematics, Tokyo, Japan
3. JENG-HUEI CHEN (陳政輝), Department of Mathematical Sciences, National Chengchi University

Analysis and Uncertainty Quantification of Digital Twins

Ying Hung

Department of Statistics, Rutgers, the State University of New Jersey, U.S.A.

yhung@stat.rutgers.edu

Abstract

We introduce a novel procedure that, given sparse data generated from a stationary deterministic nonlinear dynamical system, can characterize specific local and/or global dynamic behavior with rigorous probability guarantees. More precisely, the sparse data is used to construct a statistical surrogate model based on a Gaussian process (GP). The dynamics of the surrogate model is interrogated using combinatorial methods and characterized using algebraic topological invariants (Conley index). The GP predictive distribution provides a lower bound on the confidence that these topological invariants, and hence the characterized dynamics, apply to the unknown dynamical system (assumed to be a sample path of the GP). The focus of this paper is on explaining the ideas, thus we restrict our examples to one-dimensional systems and show how to capture the existence of fixed points, periodic orbits, connecting orbits, bistability, and chaotic dynamics.

Survival Prognostic Analysis with Copula-Graphic Estimator for Dependent Censoring

Takeshi Emura[1], Chih-Tung Yeh[2], Gen-Yih Liao[3]

[1] *Research Center for Medical and Health Data Science, the Institute of Statistical Mathematics, Tokyo, Japan*

takeshiemura@gmail.com

[2] *Department of Information Management, Chang Gung University, Taoyuan, Taiwan*

ivy202123@gmail.com

[3] *Department of Information Management, Chang Gung University, Taoyuan, Taiwan*

gyliao@gmail.com

Abstract

Survival data analysis often employs gene expressions obtained from high-throughput screening for tumor tissues from patients. When dealing with survival data, a dependent censoring phenomenon arises, and thus the traditional Cox model may not correctly identify the effect of each gene. A copula-based gene selection method can effectively adjust for dependent censoring, yielding a multi-gene predictor for survival prognosis. However, sensitivity analysis methods have not been considered to assess the impact of various types of dependent censoring on the multi-gene predictors. In this talk, we introduce a sensitivity analysis method on survival prognostic results based on a copula-graphic estimator under dependent censoring. In order to make the proposed sensitivity analysis practical, we develop an R Shiny web application. We demonstrate the proposed Shiny web application through a lung cancer dataset. This is joint work with Chih-Tung Yeh and Gen-Yih Liao from Chang Gung University.

Keywords: copula; Cox regression; dependent censoring; gene expression; lung cancer; prognostic prediction; survival prediction.

A Big-Data-Based Model of Chronic Kidney Disease and Its Applications

Jeng-Huei Chen[1], Ming-Yen Lin[2], Yi-Wen Chiu[2], Yu-Hsuan Lin[3], Yihuang Kang[4,5], Ping-Hsun Wu[2], Hsing Luh[1], Chih-Cheng Hsu[6], Shang-Jyh Hwang[2,6,7] and on behalf of the iH[3] Research Group

[1]*Department of Mathematical Sciences, National Chengchi University*

jhchen@nccu.edu.tw (J.-H.C.); slu@nccu.edu.tw (H.L.)

[2]*Division of Nephrology, Department of Internal Medicine, Kaohsiung Medical University Hospital, Kaohsiung Medical University*

1080421@kmuh.org.tw (M.-Y.L.); chiuyiwen@kmu.edu.tw (Y.-W.C.);

970392kmuh@gmail.com (P.-H.W.); sjhwang@kmu.edu.tw (S.-J. H.)

[3]*Taiwan Instrument Research Institute, National Applied Research Laboratories*

marklin@narlabs.org.tw (Y.-H. L.)

[4]*Department of Information and Management, National Sun Yat-Sen University*

ykang@mis.nsysu.edu.tw (Y.K.)

[5]*Department of Healthcare Administration and Medical Informatics, Kaohsiung Medical University*

[6]*Institute of Population Health Sciences, National Health Research Institutes, Zhunan Town*

cch@nhri.edu.tw (C.-C. H.)

[7] *School of Medicine, College of Medicine, Kaohsiung Medical University*

Abstract

Kidneys are made up of nephrons. Each nephron includes both glomerular and tubule, which filter blood and remove wastes therein. When kidneys are continually injured for various causes, their function will gradually decline. The process for such a gradual loss of kidney function is generally recognized as chronic kidney disease (CKD). In the late stage of this disease, patients might need kidney replacement therapy, such as hemodialysis. It not only greatly reduces patients' quality of life but also leads to a heavy economic burden to society. Especially, according to statistics, Taiwan has the highest incidence and prevalence of end-stage kidney disease (ESKD) worldwide. Therefore, effective prevention and control of CKD to avoid ESKD are of great importance. To achieve the goal of effective prevention and control, exploration of what causes the acceleration of kidney function loss is essential. In the literature, it is pointed out that biopsy information on CKD is fairly limited. With scarce biopsy cases, important factors leading to the accelerated kidney function progression might not be easily identified. Available longitudinal data with careful analysis could be an alternative approach to bringing insightful information. Meanwhile, the results from history data analysis can be used to predict possible progression, which facilitates CKD management. These motives motivated us to develop a data-based model for researching these related problems.

This talk presents a CKD progression model developed with data from one health surveillance program and two CKD care programs launched in Taiwan. The proposed model is Markov in nature. In addition to death and dialysis, eighteen CKD states are defined according to their estimated glomerular filtration rate (eGFR) and protein urine test results. Transition probabilities among these states are estimated with longitudinal data. Patient characteristics such as gender, smoking, drinking, exercise, diabetes, heart disease, hypertension, hyperlipidemia, etc., are also integrated into the model through regression methods, and how they affect the progression process is determined with the

collected data. With this model, probabilities of different trajectories of disease progression with definite patient characteristics may be evaluated. Crucial factors leading to rapid progression may be observed accordingly. They provide clues to what could be important issues in slowing down the progression. Meanwhile, comparing the progression processes simulated as patients in both health surveillance programs and care programs allows us to assess whether the currently launched care programs are effective. Furthermore, as it is stated, patients can also benefit from the prediction information offered by the model for their own CKD management. The model is expected to play an important role in preventing and controlling CKD progression.

Funding support statement: The National Health Research Institutes, Taiwan (grant number: NHRI-EX113-11208PI) supported the study.

Keywords: Chronic kidney disease; estimated glomerular filtration rate; Markov model.

Session V-4: Biostatistics
July 10, 2024 at 13:00-14:30
Yi-Xian Building 101 Conference Hall (050101)

◆ **Chair:**

Wen-Han Hwang (黃文瀚), Institute of Statistics, National Tsing Hua University

◆ **Speakers:**

1. George C. Tseng (曾建城), Department of Biostatistics, University of Pittsburgh, U.S.A.
2. Feng-Chang Lin (林逢章), Department of Biostatistics, University of North Carolina at Chapel Hill, U.S.A.
3. Ming-Yueh Huang (黃名鉞), Institute of Statistical Sciences, Academia Sinica

Outcome-Guided Disease Subtyping by Generative Model and Weighted Joint Likelihood in Omics Applications

George C. Tseng

Department of Biostatistics, University of Pittsburgh, U.S.A.

ctseng@pitt.edu

Abstract

With advances in high-throughput technology, molecular disease subtyping by high-dimensional omics data has been recognized as an effective approach for identifying subtypes of complex diseases with distinct disease mechanisms and prognoses. Conventional cluster analysis takes omics data as input and generates patient clusters with similar gene expression pattern. The omics data, however, usually contain multi-faceted cluster structures that can be defined by different sets of gene. If the gene set associated with irrelevant clinical variables (e.g., sex or age) dominates the clustering process, the resulting clusters may miss to capture clinically meaningful disease subtypes. This motivates the development of a clustering framework with guidance from a pre-specified disease outcome, such as lung function measurement or survival, in this paper. We propose two disease subtyping methods by omics data with outcome guidance using a generative model or a weighted joint likelihood. Both methods connect an outcome association model and a disease subtyping model by a latent variable of cluster labels. Compared to the generative model, weighted joint likelihood contains a data-driven weight parameter to balance the likelihood contributions from outcome association and gene cluster separation, which improves generalizability in independent validation but requires heavier computing. Extensive simulations and two real applications in lung disease and triple-negative breast cancer demonstrate superior disease subtyping performance of the outcome-guided clustering methods in terms of disease subtyping accuracy, gene selection and outcome association. Unlike existing clustering methods, the outcome-guided disease subtyping framework creates a new precision medicine paradigm to directly identify patient subgroups with clinical association.

Maximum Likelihood Estimation of the Silent Hypnozoite Carriage in a Malaria Randomized Clinical Trial

Feng-Chang Lin

Department of Biostatistics, University of North Carolina at Chapel Hill, U.S.A.

flin@bios.unc.edu

Abstract

Among the five species that cause malaria in humans, *Plasmodium vivax* and *Plasmodium ovale* are characterized by their silent hepatic hypnozoites that are undetectable in the blood but can reactivate and cause recurrent infection (relapse). The asymptomatic and persistent hypnozoites increase the malaria burden and hinder malaria elimination globally. This study proposes novel maximum likelihood estimators for the likelihood of relapse due to hepatic hypnozoites and silent hypnozoite carriage of individuals with no parasites in the blood. We define silent hypnozoite carriage as the possibility of carrying a hypnozoite reservoir undetectable by a blood sample. Using the maximum likelihood estimators, we also explored the risk factors associated with recurrent infection from a new mosquito bite (reinfection) that differs from the original infection. We validated our estimators under various simulation scenarios using the Newton-Raphson method and expectation-maximization (EM) algorithms to maximize the likelihood function. Lastly, we demonstrated our methods with a randomized clinical trial in Papua New Guinea. Our data analysis result showed a high incidence of *Plasmodium vivax* relapse and a high prevalence of silent *Plasmodium vivax* hypnozoite carriage. We estimated that 50.5% of the recurrent *Plasmodium vivax* infections are due to relapse (95% CI: 41.3% to 59.7%), and 64.2% of *Plasmodium vivax* patients with negative blood parasites may still have hepatic hypnozoites (95% CI: 45.4% to 82.9%). The estimation for the recurrent *Plasmodium ovale* infections is less evident due to a small sample size.

Improved Estimation Under Proportional Rates Model for Recurrent Events Data

Ming-Yueh Huang

Institute of Statistical Sciences, Academia Sinica

myh0728@stat.sinica.edu.tw

Abstract

The pseudo-partial likelihood method, known for its robustness, marginal interpretations, and ease of implementation, has become the default method for analyzing recurrent event data using Cox-type proportional rate models, as introduced in the seminal papers by Pepe & Cai (1993) and Lin et al. (2000). However, the pseudo-partial score function's construction does not account for dependency among recurrent events, leading to potential inefficiency. In this study, we explore the asymptotic efficiency of weighted pseudo-partial likelihood estimation, demonstrating that the optimal weight function depends on the unknown variance-covariance process of the recurrent event process and may lack a closed-form expression. Therefore, we propose combining a set of pre-specified weighted pseudo-partial score equations using the generalized method of moments and empirical likelihood estimation, rather than determining optimal weights. Our findings indicate that significant efficiency improvements can be readily achieved without introducing additional model assumptions. Furthermore, the proposed estimation methods can be executed using existing software. Both theoretical and numerical analyses reveal that the empirical likelihood estimator is more desirable than the generalized method of moments estimator when the sample size is sufficiently large. We present an analysis of readmission risk in colorectal cancer patients to exemplify the application of the proposed methodology. This is a joint work with Prof. Chiung-Yu Huang in University of California, San Francisco.

Parallel Session VI

Session VI-1: Biostatistics
July 10, 2024 at 14:50-16:20
E. Sun Hall

◆ **Chair:**

Ly-Yu D Liu (劉力瑜), Department of Agronomy, National Taiwan University

◆ **Speakers:**

1. Ting, Naitee (丁迺迪), Biostatistics and Data Sciences, Boehringer Ingelheim Pharmaceuticals, Inc. , U.S.A.
2. Wan Yuo Guo (郭萬祐), Taipei Veterans General Hospital and School of Medicine, National Yang Ming Chiao Tung University
3. Il Do Ha, Dept. of Statistics & data Science, Pukyong National University, Busan, South Korea

New Drug Development and Dose Finding

Naitee Ting[1]

[1] *Boehringer Ingelheim Pharmaceuticals, Inc., U.S.A.*

Naitee.ting@boehringer-ingelheim.com

Abstract

Finding the right dose(s) is one of the most important objectives in new drug development. In Phase I clinical development, one of the objectives is to escalate test doses from low to high. The low doses should be safe, then escalate up to the maximally tolerable dose (MTD). Phase II clinical trials then lower test doses to the minimal efficacious dose (MinED). Dose range of a study drug can be thought of as the doses between MinED and MTD. From this dose range, one or a few doses are selected for Phase III confirmation.

In practice, dose finding is a very difficult challenge in every phase of clinical development for new drugs. The first book "Dose finding in drug development", edited by Naitee Ting, was published in 2006. There have been lots of advancements since the publication of this book. This important field of research was pioneered by Naitee Ting.

Keywords: Dose Finding; Drug Development; Phase II Clinical Trials

The Journey of Imaging AI: From Data Governance to Clinical Implementation

Wan-Yuo Guo

Taipei Veterans General Hospital and School of Medicine, National Yang Ming Chiao Tung University

wyguo@vghtpe.gov.tw

Abstract

In this wave of artificial intelligence (AI) sweeping across all sectors worldwide, none of us can escape the impact that AI has on our work, lives, and professional practice and development. This is especially true for medical field with labor shortages, such as radiology, where AI's influence has garnered considerable attention and resources to drive short-, medium-, and long-term research and applications.

As for whether AI's impact on the field of imaging and radiology is more beneficial or detrimental, opinions vary. However, one thing is certain: healthcare providers, medical personnel, patients and their families, and the insurance systems that cover medical expenses all need to make various adjustments and changes to adapt to the era of AI in healthcare. As to whether AI's influence on the realm of medical imaging heralds a boon or a bane, perspectives indeed diverge. Yet, what remains unequivocal is the imperative for healthcare establishments, medical practitioners, patients and their families, as well as insurance frameworks linked to medical expenditure, to undertake a spectrum of adjustments and reforms. These measures are indispensable to align with the imminent era of AI in the realm of healthcare.

The field of radiology has benefited from comprehensive and rapid digitalization and the establishment of Picture Archiving and Communication System, PACS, before the end of last century. PACS is currently a fundamental infrastructure within healthcare institutions. By which medical images are archive, managed, transferred, and rederived by users at any time, from anywhere, and for any purpose. It is the one that healthcare institutions rely upon constantly and is indispensable every day. In this wave of artificial intelligence (AI) sweeping across the globe, the global radiology community and countless AI researchers are working intensively through interdisciplinary collaboration to develop AI applications in medical imaging. This represents a typical example of AI's development and application within the medical professional domain. In the sense of interdisciplinary collaboration "Data Governance and Clinical implementation" are key steps in the development of AI applications in imaging. Radiologists specialize in diagnosing diseases and detecting and assessing lesions. They are obligated to take responsibility for medical imaging data, as their professional expertise equips them to understand the processes involved in image acquisition, usage of imaging and quality control. Once AI models are developed, the critical question is how to implement them in clinical settings to assist physicians in accelerating and enhancing healthcare services and quality. This is the final stretch toward achieving the successful integration of AI in medical imaging service. The concept of "beginning with the end in mind" and "ending with the

beginning" accurately reflects the best strategic approach for the AI-driven evolution of medical imaging.

Deep Neural Networks for Frailty Models with Clustered Survival Data

Il Do Ha

Department of Statistics & data Science, Pukyong National University, Busan, South Korea

idha1353@pknu.ac.kr

Abstract

For prediction of clustered time-to-event data, we propose a new deep neural network-based semiparametric frailty model (DNN-FM). An advantage of the proposed model is that the joint maximization of the new h-likelihood provides maximum likelihood estimators (MLEs) for fixed parameters and best unbiased predictors (BUPs) for random frailties. Thus, the proposed DNN-FM is trained by using a negative profiled h-likelihood as a loss function, constructed by profiling out the non-parametric baseline hazard. Simulation studies show that the proposed method enhances the prediction performance of the existing methods (e.g. DNN based Cox model) and provides the feature selection using the multi-head attention. A real data analysis shows that the inclusion of subject-specific frailties to the DNN-Cox model helps to improve the risk prediction of the DNN based Cox model.

Keywords: Deep neural network, Frailty models, H-likelihood, Random effect.

Session VI-2: Statistics and Data Science
July 10, 2024 at 14:50-16:20
Room 210, the 2nd floor of College of Commerce (260210)

◆ **Chair:**

Nan-Cheng Su (蘇南誠), Department of Statistics, National Taipei University

◆ **Speakers:**

1. Andrei Volodin, Department of Mathematics & Statistics, University of Regina, Canada
2. Xiao Wang (王嘯), Department of Statistics, College of Science, Purdue University, U.S.A.
3. Pei-Ting Chou (周珮婷), Department of Statistics, National Chengchi University

Statistical Inference for the Ratio of Medians of Two Lognormal Distributions

Su-Fen Yang, Yu-Wei Chang, and Andrei Volodin

Department of Statistics, National Chengchi University, Taiwan

yang@mail2.nccu.tw

ychang@nccu.edu.tw

Department of Mathematics & Statistics, University of Regina, Canada

Andrei.Volodin@uregina.ca

Abstract

The lognormal distribution is used extensively to describe the distribution of positive random variables and also frequently utilized in many applications. The main purpose of our research is to establish statistical inference procedures for the ratio of means and medians of two dependent lognormal distributions based on the normal approximation and investigate their performance. Simulations are conducted to compare the performance of all methods for different values of parameters of the lognormal distributions and dependence structures. The PM2.5 data from two districts in Thailand is used to confirm the effectiveness of the proposed methods.

Keywords: Lognormal distribution, dependence structures, ratio of lognormal medians, normal approximation, statistical inference procedures.

Efficient Multi-modal Sampling via Tempered Distribution Flow

Xiao Wang

Department of Statistics, College of Science, Purdue University, U.S.A.

wangxiao@purdue.edu

Abstract

Sampling from high-dimensional distributions is a fundamental problem in statistical research and practice, and has become a central task in Bayesian computing, Monte Carlo simulation, and energy-based models. However, great challenges emerge when the target density function is unnormalized and contains multiple modes that are isolated with each other. We tackle this difficulty by fitting an invertible transformation mapping applied to the target distribution, such that the original distribution is warped into a new one that is much easier to sample from. The transformation mapping is constructed based on the normalizing ow model in deep learning. To address the multi-modality issue, our method adaptively learns a sequence of tempered distributions, which we term as a tempered distribution flow, to progressively approach the original distribution. Various experiments demonstrate the superior performance of this novel sampler compared to traditional methods. This is a joint work with Yixuan Qiu.

Unveiling the Power of Dimension Reduction in Neural Networks

Pei-Ting Chou*、Chi-Kang Wang

Department of Statistics, National Chengchi University

eptchou@nccu.edu.tw

112354032@nccu.edu.tw

Abstract

This study introduces the Siamese Fraternal Neural Network (SFNN), a groundbreaking approach merging Siamese Neural Network (SNN) principles with Principal Component Analysis (PCA) techniques. SFNN prioritizes learning data pair similarities via SNN while ingeniously integrating PCA for dimensionality reduction within the hidden layers. This innovative fusion aims to bolster computational efficiency and mitigate computational costs. The study will evaluate SFNN's predictive performance on both structured and unstructured data, exploring its potential applications. By comparing and analyzing SFNN against traditional SNN models, the research aims to demonstrate the superiority of the proposed method. Anticipated outcomes include paving new paths in statistical learning and facilitating practical implementations.

Session VI-3: Statistical Process Monitoring
July 10, 2024 at 14:50-16:20
Room 202, the 2nd floor of College of Commerce (260202)

◆ **Chair:**

Chien-Yu Peng (彭健育), Institute of Statistical Sciences, Academia Sinica

◆ **Speakers:**

1. Arthur B. Yeh (葉百堯), Department of Applied Statistics and Operations Research, Bowling Green State University, U.S.A.
2. Wei-Hang Huang (黃偉恆), Department of Statistics, Feng Chia University
3. Ming-Che Lu (呂明哲), Department of Accounting, Chaoyang University of Technology

On Monitoring and Post-Detection Diagnostics of Correlated Quality Variables of Different Types

Arthur B. Yeh,

Department of Applied Statistics and Operations Research, Schmidthorst College of Business, Bowling Green State University, Bowling Green, Ohio, U.S.A.

byeh@bgsu.edu

Abstract

Quality control applications in modern era, especially for non-manufacturing processes, often involve having to monitor correlated variables of different types, continuous, count and categorical. Most of the existing multivariate control charts implicitly assume that the correlated variables to be monitored are of the same type. Another equally challenging task in multivariate quality control which has received relatively little attention is identifying parameters that are actually out of control, when an out-of-control signal is detected on a control chart. In this talk, we will discuss how these two challenges present a unique opportunity to develop multivariate control charts which not only can monitor correlated variables of different types, but also can provide instantaneous diagnostics of out-of-control parameters. The talk will focus on discussing 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.

The Performance of S Control Charts for the Lognormal Distribution with Estimated Parameters

Wei-Heng Huang

Department of Statistics, Feng Chia University

weihhuang@mail.fcu.edu.tw

Abstract

Control charts, one of the powerful tools in statistical process control (SPC), are widely used to monitor and detect out-of-control processes in the manufacturing industry. Many researchers have pointed out the effects of using estimated parameters on the average run length (ARL) performance metric. Most of the previous literature has studied the expected value of the average run length (AARL) and the standard deviation of the average run length (SDARL) to evaluate the performance of control charts. In this article, we study the performance of three S control charts, the Shewhart S-chart, the median absolute deviation (MAD) control chart, and the lognormal S control chart, for a lognormal distribution in terms of the AARL and SDARL. Simulation results indicate the sample size to reach the specified in-control ARL value is very large. The lognormal S control chart has a smaller SDARL value than the other two S-charts. A real example is used to demonstrate how the proposed chart can be applied in practice.

Keywords: Average Run Length; average of ARL; Lognormal Distribution; Standard Deviation of ARL; S-chart

Pollution Concentration Monitoring Using a New Birnbaum-Saunders Control Chart

Ming-Che Lu[1], Su-Fen Yang[2]

[1]*Department of Accounting, Chaoyang University of Technology*

mclu@cyut.edu.tw

[2]*Department of Statistics, National Chengchi University*

yang@mail2.nccu.tw

Abstract

Air pollution monitoring is an important issue in environmental science. The Birnbaum-Saunders distribution, originally applied to describe product failure time distribution to fatigue failures and general random wear failures, is also well to describe the pollutant concentration data due to accumulations of various pollutants at the atmosphere over time. Sulfur dioxide (SO_2) is a critical factor of air pollution. Hence, it is important to monitor its concentration variation for air pollution prevention. Due to the complexity of its distribution form, there is no reliable and easy-to-use control chart for monitoring pollutant concentrations based on the Birnbaum-Saunders distribution. We found that the SO_2 concentration data follows the Birnbaum-Saunders distribution. In this study, we propose a new median control chart based on the exact sampling distribution of the monitoring statistic to detect shifts in the median of the Birnbaum-Saunders distribution. Thus, given the false alarm rate, the control limits for such control charts can be obtained precisely satisfying a preset in-control average run length using Monte Carlo simulations. The out-of-control average run lengths are calculated by simulation to evaluate the detection performance of the proposed chart when the median shifts occur. We further compare the detection performance of the proposed chart and those of the existing control charts based on asymptotic sampling distributions. In order to improve the detection ability of the proposed chart for small median shifts, an exponentially weighted moving average (EWMA) control chart is constructed. The results of numerical analyses demonstrated that the proposed EWMA chart performs much better than all existing control charts for monitoring the median of Birnbaum-Saunders distribution. Finally, the proposed control charts are applied to monitor the median of SO_2 concentrations for air pollution control, showing that both charts can effectively detect a shift in the median of SO_2 concentrations. The proposed EWMA control chart even detects out a small shift in the median of SO_2 concentrations. The results provide a continuous monitoring solution for air pollution prevention.

Keywords: air pollution monitoring, exponentially weighted moving average control chart, fatigue life, statistical process control

Session VI-4: Modern Statistics and Applications
July 10, 2024 at 14:50-16:20
Yi-Xian Building 101 Conference Hall (050101)

◆ **Chair:**

Tsai, Pi-Wen (蔡碧紋), Department of Mathematics, National Taiwan Normal University

◆ **Speakers:**

1. Chih-Li Sung (宋治立), Department of Statistics and Probability, Michigan State University, U.S.A.
2. Frederick Kin Hing Phoa (潘建興), Institute of Statistical Science, Academia Sinica
3. Yu-Wei Chang (張育瑋), Department of Statistics, National Chengchi University

Stacking Designs: Designing Multi-Fidelity Computer Experiments with Target Predictive Accuracy

Chih-Li Sung[1], Yi Ji[2], Simon Mak[3], Wenjia Wang[4], Tao Tang[5]

[1] *Department of Statistics and Probability, Michigan State University, U.S.A.*

sungchih@msu.edu

[2] *Department of Statistical Science, Duke University, U.S.A.*

yi.ji@duke.edu.

[3] *Department of Statistical Science, Duke University, U.S.A.*

sm769@duke.edu)

[4] *Data Science and Analytics Thrust, Information Hub, Hong Kong University of Science and Technology (Guangzhou)*

wenjiawang@ust.hk

[5] *Department of Mathematics, Duke University, U.S.A.*

tao.tang250@duke.edu

Abstract

In an era where scientific experiments can be very costly, multi-fidelity emulators provide a useful tool for cost-efficient predictive scientific computing. For scientific applications, the experimenter is often limited by a tight computational budget, and thus wishes to (i) maximize predictive power of the multi-fidelity emulator via a careful design of experiments, and (ii) ensure this model achieves a desired error tolerance with some notion of confidence. Existing design methods, however, do not jointly tackle objectives (i) and (ii). We propose a novel stacking design approach that addresses both goals. A multi-level reproducing kernel Hilbert space (RKHS) interpolator is first introduced to build the emulator, under which our stacking design provides a sequential approach for designing multi-fidelity runs such that a desired prediction error of $\epsilon>0$ is met under regularity assumptions. We then prove a novel cost complexity theorem that, under this multi-level interpolator, establishes a bound on the computation cost (for training data simulation) needed to achieve a prediction bound of ϵ. This result provides novel insights on conditions under which the proposed multi-fidelity approach improves upon a conventional RKHS interpolator which relies on a single fidelity level. Finally, we demonstrate the effectiveness of stacking designs in a suite of simulation experiments and an application to finite element analysis.

Keywords: Computer Experiments, Experimental Design, Finite Element Analysis, RKHS interpolator, Multilevel Modeling, Uncertainty Quantification.

An Efficient Approach for Identifying Important Biomarkers for Biomedical Diagnosis

Jing-Wen Huang1, Yan-Hong Chen2, Frederick Kin Hing Phoa2, Yan-Han Lin3, Shau-Ping Lin3

1Institute of Statistics, National Tsing Hua University

2Institute of Statistical Science, Academia Sinica

fredphoa@stat.sinica.edu.tw

3Institute of Biotechnology, National Taiwan University

Abstract

In this work, we explore the challenges associated with biomarker identification for diagnosis purpose in biomedical experiments, and propose a novel approach to handle the above challenging scenario via the generalization of the Dantzig selector. To improve the efficiency of the regularization method, we introduce a transformation from an inherent nonlinear programming due to its nonlinear link function into a linear programming framework. We illustrate the use of our method on an experiment with binary response, showing superior performance on biomarker identification studies when compared to their conventional analysis. Our proposed method does not merely serve as a variable/biomarker selection tool, its ranking of variable importance provides valuable reference information for practitioners to reach informed decisions regarding the prioritization of factors for further investigations.

Keywords: Factor Screening, Regularization Methods, Dantzig Selector, Linear Programming, Binary Responses, Biomarker Identification.

Bayesian Inference with Spike-and-Slab Priors for Differential Item Functioning Detection in a Multiple-Group IRT Tree Model

Yu-Wei Chang[1], Cheng-Xin Yang[2]

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

ychang@nccu.edu.tw

[2] *Department of Statistics, National Chengchi University*

Abstract

Group differences have practical implications in analysing data from achievement tests or questionnaires. For example, whether two persons from different demographic groups, such as gender or race, with the same shopping preferences have different shopping habits on one aspect helps store managers better design their displays. Shopping habits and shopping preferences can be measured, respectively, by items and some latent factor in a questionnaire, and the different shopping habits observed on an item are called differential item functioning (DIF). In the current study, we develop a model that accounts for between-group differences, DIF, latent factors, and missing item response data simultaneously by expanding a one-group item response tree model into a multiple-group model. Different from most of the present DIF studies where one has to iteratively select anchor items and detect DIF items, we achieve DIF detection and parameter estimation simultaneously by properly reparameterizing model parameters and applying some spike-and-slab priors (Ishwaran and Rao 2005a; Rockova and George 2018) in Bayesian estimation. Simulation studies are conducted to illustrate the validation of the proposed estimation procedure and the efficiency of DIF detection. The proposed method is further applied to a real dataset for illustration.

Keywords: Item Response Theory model, missing data, Bayesian methods, psychometrics, differential item functioning, spike-and-slab priors

Co-organizer / Sponsors

NSTC Department of International Cooperation and Science Education
(國科會國合處)

SPEC Division of Mathematics
(科學推展中心數學組)

NCCU & College of Commerce
(國立政治大學&商學院)

National Yang Ming Chiao Tung University
(國立陽明交通大學)

Department of International and Cross-strait Education
(教育部國際及兩岸教育司)

Chinese Statistical Association (Taiwan)
(中國統計學社)

Institute of Statistical Science, Academia Sinica
(中央研究院統計科學研究所)

The Chinese Institute of Probability and Statistics
(中華機率統計學會)

NCCU Risk and Insurance Research Center
(政大風險與保險研究中心)

Department of Mathematical Sciences, NCCU
(國立政治大學應用數學系)

TransGlobe Life Insurance Inc.
(全球人壽保險股份有限公司)