



# 2023 ICSA China Conference

Chengdu, Sichuan, China

June 30 - July 3, 2023



International Chinese Statistical Association

泛華統計協會

# **International Chinese Statistical Association**

## **China Conference**

**2023**

### **CONFERENCE INFORMATION, PROGRAM AND ABSTRACTS**

June 30 - July 3, 2023

Southwest Jiaotong University

Chengdu, Sichuan, China

Organized by

International Chinese Statistical Association

©2023

International Chinese Statistical Association

# Contents

|   |    |
|---|----|
| Welcome . . . . .   | 1  |
| Conference Information . . . . .  | 2  |
| ICSA Core Members . . . . .   | 2  |
| Conference Committees . . . . .   | 6  |
| Southwest Jiaotong University . . . . .   | 11 |
| School of Mathematics, Southwest Jiaotong University . . . . .                                | 13 |
| Sponsor Information . . . . .   | 15 |
| Springer Book . . . . .   | 17 |
| Journal of Nonparametric Statistics Special Issue . . . . .                                   | 19 |
| Transportation . . . . .  | 20 |
| Accommodation . . . . .   | 23 |
| Venue . . . . .   | 24 |
| Banquet . . . . .   | 30 |
| Volunteers . . . . .  | 31 |
| Program Overview . . . . .  | 32 |
| Keynote Lectures . . . . .  | 33 |
| Banquet Talk . . . . .  | 35 |
| Junior Researcher Awards . . . . .  | 37 |
| Poster Session . . . . .  | 38 |
| Scientific Program . . . . .  | 40 |
| July 1, 8:00-9:30 . . . . .   | 40 |
| July 1, 10:00-11:40 . . . . .   | 40 |
| July 1, 13:00-14:40 . . . . .   | 43 |
| July 1, 15:00-16:40 . . . . .   | 47 |
| July 1, 17:00-18:40 . . . . .   | 51 |
| July 2, 8:30-9:30 . . . . .   | 55 |
| July 2, 10:00-11:40 . . . . .   | 55 |
| July 2, 13:00-14:40 . . . . .   | 58 |
| July 2, 15:00-16:40 . . . . .   | 62 |
| July 2, 17:00-18:40 . . . . .   | 65 |
| July 3, 8:30-10:10 . . . . .  | 68 |
| July 3, 10:30-12:10 . . . . .   | 72 |
| Abstracts . . . . .   | 76 |
| Session 23CHIKT1: Keynote Lecture 1 . . . . .   | 76 |
| Session 23CHI100: Challenges in the Analysis of Survival Data with Complex Features . . . . . | 76 |
| Session 23CHI115: Recent Advanced of Modeling to Complex Censored Data . . . . .              | 77 |
| Session 23CHI118: Experimental Designs and their Applications . . . . .                       | 77 |
| Session 23CHI136: New Statistical Learning for High-Dimensional Data . . . . .                | 78 |
| Session 23CHI165: Recent Advances in Scalable Bayesian Inference . . . . .                    | 79 |
| Session 23CHI17: Statistical Methods for Complex Longitudinal and Survival Data . . . . .     | 80 |
| Session 23CHI2: Recent Developments in Network Analysis and High Dimensional Data . . . . .   | 80 |
| Session 23CHI4: Hot and Special Biostatistical Considerations in HA Guidelines . . . . .      | 81 |
| Session 23CHI59: Recent Advances in Survey Statistics and Data Science . . . . .              | 82 |
| Session 23CHI62: Recent Developments in Change Point Analysis and Related Topics . . . . .    | 82 |
| Session 23CHI68: Mendelian Randomization: Causal Inference and Beyond . . . . .               | 83 |

|   |     |
|---|-----|
| Session 23CHI69: Recent Advances in Adaptive Clinical Trial Design . . . . .  | 84  |
| Session 23CHI76: Statistical Methods and Application . . . . .  | 84  |
| Session 23CHI81: Statistical Inferences for Complex Data . . . . .  | 85  |
| Session 23CHI88: Intelligent Algorithms and Data Mining . . . . .   | 86  |
| Session 23CHI129: Optimal Designs . . . . .   | 86  |
| Session 23CHI131: Advances in Bioinformatics, Data Science, and Clinical Trial Design . . . . .                                       | 86  |
| Session 23CHI134: Message Passing and Differential Privacy . . . . .  | 87  |
| Session 23CHI144: Applications of Bayesian Methods in Educational Statistics . . . . .  | 88  |
| Session 23CHI155: Statistical Frontiers in Spatial, Single-Cell, and Single-Molecule Multi-Omics Data . . . . .                       | 89  |
| Session 23CHI166: Recent Developments in Recurrent Event Data and Panel Count Data Analysis . . . . .                                 | 89  |
| Session 23CHI26: Current Topics in Biostatistics I . . . . .  | 90  |
| Session 23CHI34: Optimization in Statistics . . . . .   | 91  |
| Session 23CHI44: New Developments in Large-Scale and High-Dimensional Inference . . . . .   | 92  |
| Session 23CHI49: New Development of Statistical Methods for Genomic, Epigenomic, and Microbiome Data . . . . .                        | 92  |
| Session 23CHI6: Case Studies in Clinical Trial Design, Analysis and Result Interpretation . . . . .                                   | 93  |
| Session 23CHI71: Addressing Challenges in Time-to-Event Data . . . . .  | 93  |
| Session 23CHI72: How to Dissect and Understand Diverse High-Throughput Data by Novel Statistical and Computational Methods? . . . . . | 94  |
| Session 23CHI74: Recent Advances in Functional Data Analysis . . . . .  | 95  |
| Session 23CHI75: Recent Statistical Advances for Complex Multi-Omics Data Analysis . . . . .  | 96  |
| Session 23CHI80: Some Advances in Analysis of High-Dimensional Complex Data . . . . .   | 97  |
| Session 23CHI83: New Advances in Statistical Methods for Health-Related Data . . . . .  | 97  |
| Session 23CHI95: Modern Statistical Process Control and Change-Point Problems I . . . . .   | 98  |
| Session 23CHI107: New Developments of Statistical Methods in Bio-Medical Studies . . . . .  | 99  |
| Session 23CHI111: Recent Development on High-Dimensional Data Analysis . . . . .  | 100 |
| Session 23CHI114: Topics on Statistical Study and Method . . . . .  | 100 |
| Session 23CHI15: Lifetime Data Analysis . . . . .   | 101 |
| Session 23CHI153: Advances in Regression Methods and Study Design . . . . .   | 101 |
| Session 23CHI169: Statistical Methods and Analysis for the Digital Asset Economy . . . . .  | 102 |
| Session 23CHI47: Advanced Statistical Methods in Omics Data Analysis . . . . .  | 103 |
| Session 23CHI48: Emerging Development in Statistical Analyses for Multi-Omic Data . . . . .   | 103 |
| Session 23CHI50: Recent Advances in Data Fusion and Integration with Real-World Applications in Healthcare . . . . .                  | 104 |
| Session 23CHI51: Big Data Analysis Based on Statistics and Machine Learning . . . . .   | 105 |
| Session 23CHI55: New Advancements in Analytical Methods using Multi-Source or Multi-Trial Data . . . . .                              | 106 |
| Session 23CHI57: Emerging Problems and Solutions in Imaging Statistics . . . . .  | 107 |
| Session 23CHI7: Data Borrowing : Methodology and Application . . . . .  | 107 |
| Session 23CHI73: Recent Developments in Medical Bioinformatics . . . . .  | 108 |
| Session 23CHI77: Recent Advances in Biostatistics . . . . .   | 109 |
| Session 23CHI84: Advances in Causal Machine Learning: Challenges and Breakthrough . . . . .   | 110 |
| Session 23CHI85: Large-Scale Dependent Data Modelling and Analysis . . . . .  | 111 |
| Session 23CHI86: Modern Statistics on Complex Data . . . . .  | 112 |
| Session 23CHI101: Advances in Analysis of Genomic and High Dimensional Data Analysis . . . . .  | 113 |
| Session 23CHI102: Frontiers in -Omics Data Analysis . . . . .   | 113 |
| Session 23CHI156: Mitigating Incomplete Data Biases-a Modern Take . . . . .   | 114 |
| Session 23CHI163: New Generation of Statisticians in Drug Development . . . . .   | 114 |
| Session 23CHI167: Causal Inference: From Practice to Theory . . . . .   | 114 |
| Session 23CHI174: Special Memorial Session to Celebrate Life of Professor Tze Leung Lai . . . . .                                     | 115 |
| Session 23CHI58: From Early Development to PAC: Application of Bayesian Analysis in Drug Development . . . . .                        | 116 |
| Session 23CHI60: Novel Machine Learning Methods to Advance Precision Medicine using Big Biomarker Data . . . . .                      | 116 |
| Session 23CHI63: Technical Advances on Analyzing Single Cell RNA-Seq and Spatial Transcriptomic Data . . . . .                        | 117 |
| Session 23CHI64: Model Estimation and Hypothesis Testing of High Dimensional Data . . . . .   | 118 |
| Session 23CHI87: Advances in Machine Learning with Applications . . . . .   | 119 |
| Session 23CHI90: Advances in Statistical Methodologies for Clustered Data Analysis and Clustered Randomized Trials Design . . . . .   | 119 |
| Session 23CHI91: Recent Development in Estimating Treatment Effects with Complex Medical Data . . . . .                               | 120 |
| Session 23CHI92: Advanced Methods for Analyzing Categorical Data . . . . .  | 121 |

|  |     |
|--|-----|
| Session 23CHI93: Limit Theorems and Inference for Time Series and Spatial Processes . . . . .  | 122 |
| Session 23CHI96: New Advances in Causal Inference and Missing Data . . . . .   | 123 |
| Session 23CHI98: Analysis of Complex Network and Text Data . . . . .   | 123 |
| Session 23CHI99: Advances in Bayesian Adaptive Design Methods for Drug Development . . . . .   | 123 |
| Session 23CHIKT2: Keynote Lecture 2 . . . . .  | 124 |
| Session 23CHI123: Recent Developments in Causal Inference . . . . .  | 124 |
| Session 23CHI130: Space-Filling Designs (I) . . . . .  | 125 |
| Session 23CHI135: Tree and Graphical Models . . . . .  | 126 |
| Session 23CHI161: Modelling Unstructured Data: Image, Network, Text and Beyond . . . . .   | 127 |
| Session 23CHI28: Recent Developments on Machine Learning and Precision Medicine . . . . .  | 127 |
| Session 23CHI30: Recent Developments in Biostatistics with their Applications in Cancer Genomics, Screening and Graph Modeling . . . . . | 128 |
| Session 23CHI31: New Developments in Missing Data Problemsand Related Areas . . . . .  | 129 |
| Session 23CHI40: Advanced Statistical Learning Methods for Complex Data . . . . .  | 130 |
| Session 23CHI41: Recent Development for Causal Inference and Personalized Medicine . . . . .   | 130 |
| Session 23CHI56: Statistical Modeling and Inferences for Complex Data Analysis . . . . .   | 131 |
| Session 23CHI66: Advances in Theory and Statistical Applications of Random Fields . . . . .  | 132 |
| Session 23CHI67: Recent Statistical Methodological Developments in Genomics . . . . .  | 132 |
| Session 23CHI8: Study Design and Statistical Considerations in Oncology Development: Methodology and Case Sharing . . . . .              | 133 |
| Session 23CHI97: Dealing with Missing Data: Recent Methodological Advances . . . . .   | 134 |
| Session 23CHI103: Modern Topics in Design of Experiments (DOE) . . . . .   | 134 |
| Session 23CHI104: Applications of Modern Statistical Methods for High-Dimensional and Complex Data . . . . .                             | 135 |
| Session 23CHI105: Survival Analysis and Quantile Regression . . . . .  | 136 |
| Session 23CHI106: Statistical Inference with Large Scale Data . . . . .  | 136 |
| Session 23CHI108: Joint Modeling of Multiple Types of Data in Health Studies . . . . .   | 137 |
| Session 23CHI113: Advanced Analytic and Innovation in Healthcare . . . . .   | 138 |
| Session 23CHI116: Development on Factorial Designs . . . . .   | 139 |
| Session 23CHI117: Stochastic Modeling of Complex Data . . . . .  | 140 |
| Session 23CHI119: Advanced Statistical Methods for Data Analysis . . . . .   | 140 |
| Session 23CHI122: Addressing Computational Challenges in Analyzing High-Throughput Genomic and Epigenetics Data . . . . .                | 141 |
| Session 23CHI171: Advances in Statistical Methods for Biomedical Studies . . . . .   | 142 |
| Session 23CHI32: Advances Developments in Statistical Methodology . . . . .  | 143 |
| Session 23CHI33: Recent Statistical Developments for Complex High-Dimensional Data . . . . .   | 143 |
| Session 23CHI36: Statistical Machine Learning and Complex Life Time Data Analysis . . . . .  | 144 |
| Session 23CHI37: Statistical Challenges for Analyzing Complex Data . . . . .   | 145 |
| Session 23CHI39: Recent Advances on Covariate Models and Applications . . . . .  | 146 |
| Session 23CHI46: Advances in Network and Complex Data Analysis . . . . .   | 147 |
| Session 23CHI109: Survival Analysis . . . . .  | 147 |
| Session 23CHI120: Advance in Statistical Methods for Large and Complex Data . . . . .  | 148 |
| Session 23CHI121: Modern Statistical and Machine Learning Methods for Analysis of -Omics Data . . . . .                                  | 149 |
| Session 23CHI125: Large Dimensional Random Matrix and Its Applications . . . . .   | 150 |
| Session 23CHI126: Recent Developments in Biostatistics and Beyond . . . . .  | 150 |
| Session 23CHI127: Experimental Designs and their Applications (II) . . . . .   | 151 |
| Session 23CHI128: Space-Filling Designs (II) . . . . .   | 152 |
| Session 23CHI133: Statistical and Machine Learning Methods for Biomedical Research . . . . .   | 153 |
| Session 23CHI137: Recent Developments in Machine Learning and Causal Inference . . . . .   | 153 |
| Session 23CHI138: Recent Developments in Applied Probability and Statistics . . . . .  | 154 |
| Session 23CHI139: Statistical Modeling for Neuroimaging Data . . . . .   | 155 |
| Session 23CHI140: Network Data and Large-Scale Computing . . . . .   | 156 |
| Session 23CHI170: Checking Structural Change of Complex Data . . . . .   | 157 |
| Session 23CHI25: Current Topics in Biostatistics II . . . . .  | 157 |
| Session 23CHI29: Advanced Statistical Considerations on Contemporary Clincal Trials . . . . .  | 158 |
| Session 23CHI61: Recent Methodological Advances in Survey Statistics . . . . .   | 158 |
| Session 23CHI94: Modern Statistical Process Control and Change-Point Problems II . . . . .   | 159 |

|  |     |
|--|-----|
| Session 23CHI132: Advanced Methods in Adaptive Randomization Designs . . . . .   | 160 |
| Session 23CHI14: New Developments in Nonparametric and Semiparametric Methods for Complex Data . . . . .                             | 161 |
| Session 23CHI143: Recent Developments of Statistical Methods on Missing Data with Applications . . . . .                             | 161 |
| Session 23CHI157: Statistical Advances in Biomedical Data Analysis . . . . .   | 162 |
| Session 23CHI158: Statistical Inferences in Computational Biology and Genetics . . . . .   | 163 |
| Session 23CHI18: Advanced Statistical Methods for Complex Data . . . . .   | 163 |
| Session 23CHI19: Recent Development in Extreme Value Theory . . . . .  | 164 |
| Session 23CHI20: High Dimensional Modeling and Inference . . . . .   | 165 |
| Session 23CHI21: Statistical Learning for Regression Analysis and Change Point Detection . . . . .                                   | 166 |
| Session 23CHI23: Statistics Process Control and Change-Point Analysis . . . . .  | 166 |
| Session 23CHI24: Recent Developments in the Analysis of Complex Lifetime Data . . . . .  | 167 |
| Session 23CHI27: Stochastic Modeling and Inference of Epidemiological and Industrial Data with Complex Structures . . . . .          | 168 |
| Session 23CHI79: Novel Bayesian Methods for Complex Dataand their Applications . . . . .   | 168 |
| Session 23CHI1: Sufficient Dimension Reduction and Beyond . . . . .  | 169 |
| Session 23CHI10: Novel Bayesian Adaptive Clinical Trial Designs and Methods for Precision Medicine . . . . .                         | 170 |
| Session 23CHI11: Advanced Biostatistics and Bioinformatics Approaches for Medical Research . . . . .                                 | 170 |
| Session 23CHI12: Topics on Latent Variable Models and Minorisation-Maximisation Algorithm . . . . .                                  | 171 |
| Session 23CHI13: New Developments in Statistical Detection . . . . .   | 172 |
| Session 23CHI141: Fator Modelling and Data Analysis in Large-Scale Datasets . . . . .  | 173 |
| Session 23CHI142: Recent Advances in Spatio Temporal Modelling . . . . .   | 174 |
| Session 23CHI145: Human-Centric Statistical Learning . . . . .   | 174 |
| Session 23CHI147: Statistical Methods and Applications on Unstructured Data . . . . .  | 176 |
| Session 23CHI148: New Developments on Design of Experiments and Sampling Methods . . . . .   | 176 |
| Session 23CHI162: New Methods and Applications of Reinforcement Learning for Complex Data . . . . .                                  | 177 |
| Session 23CHI164: Complex Statistical Modelling and Testing . . . . .  | 178 |
| Session 23CHI168: Recent Advances in Sequencing Data Analysis, Reinforcement Learning and Missing Data Handling . . . . .            | 179 |
| Session 23CHI3: Computational Approaches to Single-Cell Genomics and Spatial-Omics Data Analysis and Clinical Applications . . . . . | 180 |
| Session 23CHI43: Recent Advancements in Bayesian Methods for Causal Inference . . . . .  | 181 |
| Session 23CHI89: Modern Learning Methods for Causal Inference and Survey Inference . . . . .   | 181 |
| Session 23CHI9: Some Recent Developments in Deep Learning . . . . .  | 182 |
| Session 23CHI110: Recent Advances on Functional Data Analysis . . . . .  | 183 |
| Session 23CHI112: Recent Advances in Variable Selection and Regression Analysis with Interval-Censored Data                          | 184 |
| Session 23CHI124: Order-of-Addition Experimental Designs . . . . .   | 184 |
| Session 23CHI146: Recent Developments in Complex Data Structure Analysis . . . . .   | 185 |
| Session 23CHI150: Statistical Analysis of Massive Data and Network Data . . . . .  | 186 |
| Session 23CHI151: Statistical Learning for Large-Scale and Complex Data . . . . .  | 187 |
| Session 23CHI152: Causal Reinforcement Learning . . . . .  | 187 |
| Session 23CHI172: Junior Researcher Award Session . . . . .  | 188 |
| Session 23CHI35: Non-Euclidean Data Analysis . . . . .   | 189 |
| Session 23CHI38: Recent Advances in Privacy-Protected Data Collection and Analysis . . . . .   | 189 |
| Session 23CHI42: Recent Developments in Clinical Trial Design and Data Analysis . . . . .  | 190 |
| Session 23CHI45: Recent Advances in Statistical Methodology . . . . .  | 190 |
| Session 23CHI5: Innovative Statistical Methods . . . . .   | 191 |
| Session 23CHI78: Recent Developments in Survival Data Analysis . . . . .   | 192 |
| Session 23CHI82: Advances in Statistical Methods and Inference for Complex Biomedical Data . . . . .                                 | 193 |
| Index of Authors . . . . .   | 194 |

## ICSA 2023 China Conference June 30- July 3, Chengdu, Sichuan, China

Welcome to the International Chinese Statistical Association (ICSA) 2023 China Conference!

The ICSA 2023 China Conference will be held at Chengdu, Sichuan, China from June 30 – July 3, 2023. It is jointly organized by ICSA, Southwest Jiaotong University (SWJTU) and Sichuan Association of Applied Statistics. This year we gather here for the ICSA China Conference after the COVID-19 pandemic. This is the 6th annual meeting for ICSA China Conference. The theme of this conference is the "***Data Science with Applications to Big Data Analysis and AI***", which is in recognition of artificial intelligence and data science with new challenges in the big data era.

The executive and organizing committees have been working diligently to put together a very strong and comprehensive program, including two keynote lectures, 164 invited sessions, poster sessions, junior researcher award session, and exciting social events. Our scientific program includes keynote lectures from leading experts in statistics ***Dr. Jun Liu*** (Harvard University, USA) and ***Dr. Fang Yao*** (Peking University, China), numerous invited talks, 49 poster presentations reflecting recent challenges in statistics, business statistics, and biostatistics, which are closely related to recent progress in big data analysis and artificial intelligence.

With your full support, this conference attracts more than 931 statisticians and data scientists working in academia, government, and industry from all over the world. We hope that the conference provides great opportunities for learning, networking, collaborations and recruiting. You will share the thoughts and ideas with conference guests, and receive inspirations from old research ideas and develop new ones. The local organizing committee and more than 50 student volunteers led by Professor Haitao Zheng in SWJTU have made a great effort to arrange the meeting logistics and social events in this conference. The social events include the opening mixer (Friday, June 30 evening) and the banquet (Sunday, July 2 evening, banquet speaker ***Dr. Zongben Xu***). We believe this conference will be a memorable, interesting and enjoyable experience for all of you.

In addition, I am happy to announce that the ***Journal of Nonparametric Statistics*** will publish a special issue with selected papers presented in the conference. Details about how to submit the papers can be found in the conference program book.

The city of Chengdu enjoys a mild climate throughout the year, and is accessible from most cities across China. Chengdu provides numerous opportunities for tour, shopping and lodging, etc. In particular, Chengdu is well-known for lovely giant pandas, and delicious Sichuan food. Moreover, Chengdu is a very big city, with numerous natural, cultural, and historical attractions around it. It is our sincere hope you have great opportunities to experience the wonderful activities during your stay in the beautiful city of Chengdu.

**Thanks for coming to the ICSA 2023 China Conference in Chengdu!**

Yichuan Zhao, on behalf of  
ICSA 2023 China Conference Executive and Organizing Committee

## ICSA 2023 Core Members

### EXECUTIVES

President: Gang Li ([vli@ucla.edu](mailto:vli@ucla.edu))

Past-President: Zhezhen Jin ([zj7@columbia.edu](mailto:zj7@columbia.edu))

President-Elect: Xun Chen ([xun.chen@sanofi.com](mailto:xun.chen@sanofi.com))

Executive Director: Jun Zhao (2023-2025, [executive.director@icsa.org](mailto:executive.director@icsa.org))

ICSA Treasurer: Rui Feng (2022-2024, [treasurer@icsa.org](mailto:treasurer@icsa.org) )

The ICSA Office Manager: Grace Ying Li

Email: [oicsa@icsa.org](mailto:oicsa@icsa.org)

### BOARD of DIRECTORS

Shu-Hui Chang (2021-2023, [shuhui@ntu.edu.tw](mailto:shuhui@ntu.edu.tw))

Yong Chen (2021-2023, [ychen123@mail.med.upenn.edu](mailto:ychen123@mail.med.upenn.edu))

Chenlei Leng (2021-2023, [c.leng@warwick.ac.uk](mailto:c.leng@warwick.ac.uk))

Xiaodong Luo (2021-2023, [xiaodong.luo@sanofi.com](mailto:xiaodong.luo@sanofi.com))

Yang Song (2021-2023, [Yang\\_Song@vrtx.com](mailto:Yang_Song@vrtx.com))

Gang Li (2022-2024, [Gang\\_Li@eisai.com](mailto:Gang_Li@eisai.com))

Annie Qu (2022-2024, [aqu2@uci.edu](mailto:aqu2@uci.edu))

Lan Wang (2022-2024, [lanwang@mbs.miami.edu](mailto:lanwang@mbs.miami.edu))

Hao (Helen) Zhang (2022-2024, [hzhang@math.arizona.edu](mailto:hzhang@math.arizona.edu))

Xingqiu Zhao (2022-2024, [xingqiu.zhao@polyu.edu.hk](mailto:xingqiu.zhao@polyu.edu.hk))

Ming Tan (2023-2025, [mtt34@georgetown.edu](mailto:mtt34@georgetown.edu))

Huazhen Lin (2023-2025, [linhz@swufe.edu.cn](mailto:linhz@swufe.edu.cn))

Min Zhang (2023-2025, [mzhangst@umich.edu](mailto:mzhangst@umich.edu))

Li Wang (2023-2025, [li.wang1@abbvie.com](mailto:li.wang1@abbvie.com))

Yanping Wang (2023-2025, [WANG\\_YANPING@LILLY.COM](mailto:WANG_YANPING@LILLY.COM))

### COMMITTEES:

#### Program Committee

**Chair:** Xinpeng Cui (2023, [xpcui@ucr.edu](mailto:xpcui@ucr.edu))

#### Members:

Pei Wang (2021-2023, JSM Representative 2022, [pei.wang@mssm.edu](mailto:pei.wang@mssm.edu))

Jianguo (Tony) Sun (2022-2024, JSM Representative 2023, ICSA International Conference 2022, [sunj@missouri.edu](mailto:sunj@missouri.edu))

Samuel Wu (2021-2023, ICSA Symposium 2022, [samwu@biostat.ufl.edu](mailto:samwu@biostat.ufl.edu))

Gongjun Xu (2022-2024, ICSA Symposium 2023, [gongjun@umich.edu](mailto:gongjun@umich.edu))

Jian Kang (2022-2024, ICSA Symposium 2023, [jiankang@umich.edu](mailto:jiankang@umich.edu))

Xin-Yuan Song (2021-2023, ICSA International Conference 2022, [xysong@sta.cuhk.edu.hk](mailto:xysong@sta.cuhk.edu.hk) )

# ICSA Officers and Committees

Hulin Wu (2023-2025, Hulin.Wu@uth.tmc.edu)  
Lihui Zhao (2022-2024, Lihui.zhao@northwestern.edu)  
Dandan Liu (2023-2025, ICSA Symposium 2024, dandan.liu@vumc.org)  
Qingxia Chen (2023-2025, ICSA Symposium 2024, cindy.chen@vanderbilt.edu)  
Yichuan Zhao (2022-2024, ICSA China Conference 2023, yichuan@gsu.edu)

## Awards Committee

**Chair:** Zhigang Li (2023, zhigang.li@ufl.edu)

**Members:**

Chunming Zhang (2023-2025, cmzhang@stat.wisc.edu)  
Wei Wu (2023-2025, wwu@fsu.edu)  
Lu Tian (2023-2025, lutian@stanford.edu)  
Yong Chen (2023-2025, ychen123@pennmedicine.upenn.edu)  
Xuezhou Mao (2023-2025, Xuezhou.Mao@modernatx.com)  
Mingxiu Hu (2021-2023, mhu@nektar.com)  
Jianxin Shi (2021-2023, jianxin.shi@nih.gov)  
Ying Wei (2021-2023, yw2148@cumc.columbia.edu)  
Jie Peng (2021-2023, jiepeng@ucdavis.edu)  
Bo Huang (2022-2024, Bo.Huang@pfizer.com)  
Charles Ma (2022-2024, tianzhou.ma0105@gmail.com)  
Jiayang Sun (2022-2024, jsun@case.edu)

## Nominating and Election Committee

**Chair:** Yichuan Zhao (2023, yichuan@gsu.edu)

**Members:**

Wenqing He (2023-2025, whe@stats.uwo.ca)  
Yuanyuan Lin (2022-2023, ylin@sta.cuhk.edu.hk)  
Bo Fu (2022-2023, bo.fu.stat@gmail.com)  
Sijian Wang (2022-2023, sijian.wang@stat.rutgers.edu)  
Henry Horng-Shing Lu (2022-2023, hslu@stat.nycu.edu.tw)  
Hongjian Zhu (2023-2025, Hongjian.zhu@abbvie.com)  
Zhigang Li (2023-2025, zhigang.li@ufl.edu)

## Special Lecture Committee

**Chair:** Ming Tony Tan (2023, mtt34@georgetown.edu)

**Members:**

Hongzhe Lee (2023-2025, hongzhe@pennmedicine.upenn.edu)  
Aiyi Liu (2021-2023, liua@mail.nih.gov)  
Gang Li (2021-2023, Gang\_Li@eisai.com)  
Jianguo Sun (2022-2024, sunj@missouri.edu)  
Xiaonan Xue (2022-2024, Xiaonan.xue@einsteinmed.org)

## Publication Committee

**Chair:** Runze Li (2023, rzli@psu.edu)

**Members:**

Linda Zhao (2023-2025, lzhao@wharton.upenn.edu)

# ICSA Officers and Committees

Hongkai Ji (2022-2024, Co-Editors of SIB, hji@jhu.edu)  
Joan Hu (2021-2023, Co-Editors of SIB, joan\_hu@sfsu.ca )  
Rong Chen (2021-2023, Co-Editors of Statistica Sinica, rongchen@stat.rutgers.edu )  
Su-Yun Huang (2021-2023, Co-Editors of Statistica Sinica, syhuang@stat.sinica.edu.tw )  
Xiaotong Shen (2021-2023, Co-Editors of Statistica Sinica, xshen@umn.edu )  
Ding-Geng (Din) Chen (2020-2023, Editor of ICSA book series, dinchen@email.unc.edu )  
Ming Wang (2021-2023, Editor for ICSA Bulletin, mwang@phs.psu.edu )  
Jun Zhao (2023-2025, Executive Director of ICSA, executive.director@icsa.org )  
Sheng Luo (2022-2024, sheng.luo@duke.edu)  
Ying Ding (2022-2024, yingding@pitt.edu)

## Membership Committee

**Chair:** Zhigen Zhao (2023, zhaozhg@temple.edu)

**Members:**

Tu Xu (2023-2025, Tu\_Xu@vrtx.com)  
Yifei Sun (2021-2023, ys3072@cumc.columbia.edu )  
Fei Huang (2022-2024, feihuang@unsw.edu.au)  
Xun Chen (2022-2024, xun.chen@sanofi.com)  
Wei Zhang (2022-2024, wei.zhang@boehringer-ingelheim.com)  
Anru Zhang (2022-2024, anru.stat@gmail.com, anru.zhang@duke.edu)

## IT Committee

**Chair:** Chengsheng Jiang (2023, website@icsa.org)

## Archive Committee

Chair: Naitee Ting (2023, naitee.ting@boehringer-ingelheim.com)

**Members:**

Xin (Henry) Zhang (2021-2023, henry@stat.fsu.edu)  
Rui Miao (2021-2023, ruimiao@gwu.edu )  
Xin Tian (2022-2024, tianx@nhlbi.nih.gov)  
Jun Yan (2022-2024, jun.yan@uconn.edu)

## Finance Committee

**Chair:** Rui Feng (2022-2024, ruifeng@pennmedicine.upenn.edu)

**Members:**

Xin He (2021-2023, xinhe@umd.edu )  
Rochelle Fu (2022-2024, fur@ohsu.edu)

## Financial Advisory Committee

**Chair:** Fang Chen (2023, FangK.Chen@sas.com)

**Members:**

Nianjun Liu (2021-2023, liunian@indiana.edu)  
Xiangqin Cui (2021-2023, xiangqin.cui@emory.edu)  
Rochelle Fu (2022-2024, fur@ohsu.edu)  
Yuan Jiang (2022-2024, yuan.jiang@stat.oregonstate.edu)  
Hongliang Shi (2020-2023, hongliangshi15@gmail.com)

# ICSA Officers and Committees

Lingzi Lu Award Committee (ASA/ICSA)

**Chair:** Ivan Chan (2022-2024, ivan\_chan@bms.com)

Shelly Hurwitz (2017-2023, hurwitz@hms.harvard.edu)

Laura J Meyerson (2020-2023, laurameyerson@msn.com)

Kelly Zou (2022-2024, Kelly.Zou@viatris.com)

ICSA Representative to JSM Program Committee

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

Yingwen Dong, (2024, yingwen.dong@sanofi.com)

## AD HOC COMMITTEES

2023 ICSA China Conference

**Chair:** Haitao Zheng (htzheng@sjtu.edu.cn), Yichuan Zhao (yichuan@gsu.edu)

2023 Applied Statistics Symposium

**Chair:** Gongjun Xu (gongjun@umich.edu), Jian Kang (jiankang@med.umich.edu)

## CHAPTERS

### **ICSA-Canada Chapter**

Joan Hu (Chair, joan\_hu@sfu.ca)

### **ICSA-Midwest Chapter**

Xiaohong Huang (Chair, xiaohong.huang@abbvie.com)

### **ICSA-Taiwan Chapter**

Henry Horng-Shing Lu (Chair, hslu@stat.nycu.edu.tw)

### Executive Committee

Co-Chair: Dan Yang, Southwest Jiaotong University

Co-Chair: Yichuan Zhao, Georgia State University

Zhezhen Jin, Columbia University

Gang Li, University of California, Los Angeles

Mengling Liu, New York University

Hulin Wu, University of Texas Health Science Center at Houston

Jun Zhao, Antengene

Haitao Zheng, Southwest Jiaotong University

### Scientific Program Committee

Yichuan Zhao, Georgia State University, Committee Chair

Ming-Hui Chen, University of Connecticut

Nelson Chen, University of Illinois at Chicago

Somnath Datta, University of Florida

Xindong Feng, Shanghai University of Finance and Economics

Bo Fu, Astellas China

Haoda Fu, Eli Lilly

Wenqing He, Western University

Joan Hu, Simon Frazer University

Li-Shan Huang, National Tsing Hua University

Zhezhen Jin, Columbia University

Jian Kang, University of Michigan

Mei-Ling Ting Lee, University of Maryland

Gang Li, University of California, Los Angeles

Gang Li, Eisai

Yi Li, University of Michigan

Yisheng Li, The University of Texas MD Anderson Cancer Center

Huazhen Lin, Southwestern University of Finance and Economics

## Conference Committees

Aiyi Liu, NIH

Lei Liu, Washington University in St. Louis

Qi Long, University of Pennsylvania

Wenbin Lu, North Carolina State University

Xuewen Lu, University of Calgary

Ying Lu, Stanford University

Yajun Mei, Georgia Institute of Technology

Bin Nan, University of California, Irvine

Steve Qin, Emory University

Annie Qu, University of California, Irvine

Yiyuan She, Florida State University

Tony Sun, University of Missouri

Wenguang Sun, Zhejiang University

Tiejun Tong, Hong Kong Baptist University

Chunjie Wang, Changchun University of Technology

Xueqin Wang, University of Science and Technology of China

Samuel Wu, University of Florida

Yingcun Xia, National University of Singapore

Song Yang, NIH

Heping Zhang, Yale University

Jin-Ting Zhang, National University of Singapore

Hongyu Zhao, Yale University

Jun Zhao, Antengene

Haitao Zheng, Southwest Jiaotong University

Hongjian Zhu, AbbVie

## Junior Researcher Award Committee

Lily Wang, George Mason University, Committee Chair

Pengsheng Ji, University of Georgia

Zhigang Li, University of Florida

## Conference Committees

Ruiyan Luo, Georgia State University  
Yao Xie, Georgia Institute of Technology  
Yanxun Xu, Johns Hopkins University  
Ao Yuan, Georgetown University  
Jing Zhang, Georgia State University

## IT and Website Committee

Co-Chair: Chengsheng Jiang, Flatiron Health  
Co-Chair: Xiang Huang, Southwest Jiaotong University, China  
Yinghui Yang, Southwest Jiaotong University, China  
Junjie Qiu, Southwest Jiaotong University, China  
Suzi Zhang, Southwest Jiaotong University, China

## Local Organization Committee

Co-Chair: Han Yang, Southwest Jiaotong University, China  
Co-Chair: Lu Wang, Southwest Jiaotong University, China  
Co-Chair: Jiayin Tang, Southwest Jiaotong University, China  
Co-Chair: Haitao Zheng, Southwest Jiaotong University, China  
Gaoxiu Qiao ,Southwest Jiaotong University, China  
Xiang Huang ,Southwest Jiaotong University, China  
Ruixin Ren ,Southwest Jiaotong University, China  
Chunming Zhao ,Southwest Jiaotong University, China  
Suzi Zhang ,Southwest Jiaotong University, China  
Lianwen Zhao, Southwest Jiaotong University, China  
Cheng Liu,Southwest Jiaotong University, China  
Qin Wang,Southwest Jiaotong University, China  
Shijuan Cheng,Southwest Jiaotong University, China  
Yinghui Yang, Southwest Jiaotong University, China  
Baoying Yang, Southwest Jiaotong University, China

## Conference Committees

Lei Huang, Southwest Jiaotong University, China

Hongfan Zhang, Southwest Jiaotong University, China

Chongshuang Chen, Southwest Jiaotong University, China

Jianpeng Wang, Southwest Jiaotong University, China

Dailin Yuan, Southwest Jiaotong University, China

Hao Peng, Southwest Jiaotong University, China

## Fundraising Committee

Gang Li, Eisai, Committee Chair

Na Hu, Boehringer Ingelheim

Hongliang Shi, Blueprint Medicines

Haitao Zheng, Southwest Jiaotong University

## Program Book Committee

Co-Chair: Dayu Sun, Indiana University/Emory University

Co-Chair: Yuanyuan Guo, Biogen Inc.

Lu Wang, Southwest Jiaotong University

Yinghui Yang, Southwest Jiaotong University

Chengcheng Zhao, Southwest Jiaotong University

Rui Wu, Southwest Jiaotong University

Hang Ruan, Southwest Jiaotong University

Chunming Zhao, Southwest Jiaotong University

## Poster Session Committee

Co-Chair: Lu Wang, Southwest Jiaotong University, China

Co-Chair: Lei Huang, Southwest Jiaotong University, China

Wenhai Gui, Beijing Jiaotong University, China

Guanyu Hu, University of Missouri, USA

## Conference Committees

Yi Li, Southwestern University of Finance and Economics, China

Gaoxiu Qiao, Southwest Jiaotong University, China

Ruixin Ren, Southwest Jiaotong University, China

Baoying Yang, Southwest Jiaotong University, China

西南交通大学创建于1896年，前身是北洋官铁路局创办的山海关北洋铁路官学堂。1900年，由于八国联军入侵，学堂被迫停办。1905年，学堂在唐山复校，改名为唐山铁路学堂。1906年，增设矿科，更名为山海关内外路矿学堂。1908年，学堂由清政府邮传部直辖，更名为邮传部唐山路矿学堂。1912年，中华民国成立。清政府的邮传部被民国政府交通部取代，学堂归交通部直辖，学校更名为交通部唐山铁路学校。1913年，学校奉教育部、交通部令，更名为唐山工业专门学校。1921年，北洋政府交通部组建交通大学，总部设在北京，下设北京、唐山、上海三个学校。我校更名为交通大学唐山学校。1922年，交通部改组交通大学，下设唐山大学和南洋大学，学校更名为交通部唐山大学。1928年2月，北洋政府交通部指令唐山大学改名唐山交通大学。同年6月，国民政府宣布南北统一，唐山交通大学改称第二交通大学。同年，交通部重组交通大学，总部在上海。我校更名为交通大学唐山土木工程学院。同年10月，国民政府设立铁道部，我校暂归铁道部直辖。11月，交通大学移归铁道部后，交通大学改称铁道部交通大学，下设上海本部、北平铁道管理学院和唐山工程学院，学校更名为交通大学唐山工程学院。



1937年“七七事变”后，我校校园被日军占领。在全校师生、校友的努力下，学校于年底在湖南湘潭复校。1938年3月，教育部指令，交通大学北平铁道管理学院暂行并入唐山工程学院。5月，学校迁往湖南湘乡杨家滩。1938年，武汉沦陷，举校再次被迫西迁。1939年，学校在贵州平越古城（今福泉市）复课。1941年7月，教育部下令将学校校名改

为国立交通大学唐山工程学院、北平铁道管理学院。由于此校名在校内引起争议，1942年1月，教育部更改前令，组建国立交通大学贵州分校，下设唐山工程学院和北平铁道管理学院。1944年11月，日军攻占贵州独山，学校被迫再次迁校到四川璧山办学。抗日战争胜利后，1946年8月，接教育部令，学校更名为国立唐山工学院，归教育部直辖，迁回唐山原址办学。1949年，新中国成立，学校由中央军委铁道部接管，组建中国交通大学，本部在北京，下设唐山工学院和北京铁道管理学院，学校更名为中国交通大学唐山工学院。1950年8月，学校更名为北方交通大学唐山工学院。1952年，全国高等院校进行院系调整，我校采矿、冶金、化工、建筑、水利、通讯等系（组）师生调整到外校，我校更名为唐山铁道学院。1964年9月，根据中共中央建设大三线的精神，铁道部决定我校迁至四川峨眉。1972年，学校更名为西南交通大学。1989年，总校迁往成都，（峨眉成为分校，后改为校区）。2002年在成都犀浦建设新校区，遂形成今日“一校、两地、三校区”的办学格局。

1928年，国民政府成立后，交通部重组交通大学，总部在上海，我校当时名为交通大学唐山土木工程学院。1930年5月2日，交通大学公布校训：“精勤求学，敦笃励志，果毅力行，忠恕任事”。1947年5月15日，在唐山工学院（西南交通大学）建校51周年，唐山复校42周年校庆之际，我校重新确定原交通大学之校训为唐山工学院之院训。院训曰：“精勤求学，敦笃励志，果毅力行，忠恕任事”。

“俟实扬华”的由来：1916年春，教育部在北京举行全国高等学校作业成绩展览评比，我校以94分的优异成绩荣获全国第一名。同年12月，教育部为此除给我校颁发优等奖状外，还由教育总长范源濂特奖我校“俟实扬华”匾额一方，“俟”就是等待的意思；“实”则有三层意思，即果实、坚实、诚实和实事求是；“扬”有飞扬，传播之意；“华”有中华简称、华丽、美丽等意。四个字组合在一起，有等待果实成熟，即寓意培养出人才，同时使人才更加务实、诚实，从而扬我中华，振兴中华，复兴中华。“扬华”也有扬弃浮华、追求一种务实的、不尚虚华之意。

“自强不息”是学校在长期的颠沛流离的办学经历中总结出来的，该词语出自《周易》的“天行健，君子以自强不息”，原意为，天上的日月星辰是不分昼夜、永恒运动的，所以“天”是“刚健”的，人应效法天，积极进取，永不停息。对交大来说，其中有四层意思：永不停息、知难而上、完善自我和艰苦奋斗。

“俟实扬华，自强不息”的交大精神包含有四种内涵：首先是“爱国至上、振兴中华”；其次是“严谨严格、求真务实”；再次是“爱校如家、敬业奉献”；最后是“开拓创新、艰苦奋斗”

## 西南交通大学数学学院简介

西南交通大学数学学院的前身是创办于1958年的唐山铁道学院数理力学系，1985年学校正式成立数学系，2008年组建数学学院。学院以“坚持发展基础数学，建设以应用、计算、统计与数据科学为特色的西南交通大学数学学院”为学科发展思路，建设成为国内一流数学与统计人才培养和科学研究基地为发展目标。

学院拥有一支博学敬业的高水平师资队伍。学院下设数学系、统计系、信息与计算科学系、公共数学教学与研究部。现有教职工140人，其中专任教师124人，专任教师中有近90%具有博士学位，半数的教师具有海外求学或访学经历。在职教授14名、副教授58名，包括中科院院士（双聘）1名，国家青年拔尖人才1名，国家有突出贡献的中青年专家1名，四川省“千人计划专家”7名，四川省学术和技术带头人1名及后备人选13名，四川省“杰出青年基金”获得者3名，“天府万人计划·天府科技菁英人才”1名。数学学院先后为国家培养了近三千名毕业生奋斗在祖国建设的各个岗位上，其中包括我国著名数学家、全国劳动模范、国际戴维逊奖和第十三届华罗庚数学奖获得者侯振挺。

学院学科体系完整，科研平台完善，形成了本硕博三个层次完整的人才培养格局。学院现有数学一级学科博士点和统计学一级学科硕士点，数学与应用数学、统计学和数据科学与大数据技术三个本科专业（三个本科专业皆为国家一流建设专业）。现有在校本科生700余人，硕士研究生200余人，博士研究生40余人。每年本科毕业生中有近50%的学生到国内外高校进一步深造，硕士研究生中近30%选择继续攻读博士学位，近5年来还有多位研究生获国家留学基金委资助到国外哈佛等著名大学攻读博士学位。

数学学院为学生创造一流的学习环境，加强基础，结合应用，重视实践。学院建设有系统可信性自动验证国家地方联合工程实验室、四川省系统可信性自动验证工程实验室、金融大数据研究院、数学中心四个科研平台和相应科研团队。



学院积极开展科学研究。学院教师在代数、几何、泛函分析、微分方程与动力系统、符号演算与机器证明、数值计算、序列编码、不确定性信息处理、可靠性理论、金融统计与生物统计等方向从事学术研究。科研成果先后获教育部自然科学二等奖、湖北省自然科学一等奖、上海市自然科学二等奖、四川省科技进步奖二等奖和全国百篇优秀博士学位论文奖。近五年来共主持或主研各级科研项目近200项（含国家自然科学基金项目50余项），科研经费超三千万元；在国内外重要学术期刊杂志上发表论文800余篇。多名青年教师在国际顶尖期刊如《Ann Stat》、《Math Annalen》、《Compositio Mathematica》、《J Differ Equations》、《Transaction of AMS》、《Commun Math Phys》、《J Funct Anal》等发表学术论文。多名教师担任国内外多种学术刊物的编委、专业学会的理事长、副理事长。

学院学术交流活跃，教师与美国、加拿大、巴西、波兰、法国、英国、日本、新加坡、香港等地的高校学者开展学术交流频繁且富有成效。近年来主办承办多场国际或全国性学术会议，每年邀请多位国内外著名专家学者到校开设学术报告，进行学术交流。学院

的师生也积极参加国内外的学术会议，在国内外重要学术会议上报告。学院的学术交流先后得到中国科学院院士华罗庚，刘应明，陈木法，严加安，李安民，袁亚湘等的亲切指导和帮助。

数学学院秉承学校“严谨治学、严格要求”的双严传统，承担着全校所有专业的本科、硕士、博士各层次数学类课程的教育、教学工作，建设有《概率论A》、《数学建模》两门国家一流课程；《高等数学》、《线性代数》、《概率与数理统计》、《数学建模》等省级精品课程。近5年来主持国家级教改项目2项；主持省部级教改项目13项。获省部级教学成果奖10项；校级教学成果奖22项。出版教材、专著36部。

由数学学院直接负责指导、培训的全国大学生数学竞赛、大学数学数学建模竞赛、市场调查分析大赛、全国研究生建模竞赛等参赛工作屡创佳绩，获奖队数位列全国前茅，其中在2021年获得全国大学生数学建模最高奖项--“高教社杯”。

数学学院广大师生将继续秉承学校“俟实扬华，自强不息”的精神，锐意进取，迎接挑战，为把学院建设成为国内一流的有特色的数学学院而努力奋斗！



## The 2023 ICSA China Conference Sponsors

The 2023 ICSA China Conference Sponsors Committees gratefully acknowledge the generous support of our sponsors below.

### Gold Sponsors



西南交通大学  
Southwest Jiaotong University



National Science Foundation  
WHERE DISCOVERIES BEGIN

abbvie

成都市博览局

### Silver Sponsors



### Bronze Sponsors

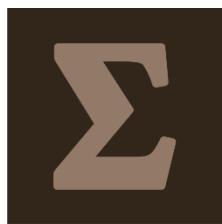


### Additional Sponsors



International  
Statistical  
Institute

# Sponsors Information



# mathematics



an Open Access Journal by MDPI

► [www.mdpi.com/journal/mathematics](http://www.mdpi.com/journal/mathematics)

*Mathematics* (ISSN 2227-7390) is a peer-reviewed, open access journal which provides an advanced forum for studies related to mathematics, and is published semimonthly online by MDPI. The European Society for Fuzzy Logic and Technology (EUSFLAT) and International Society for the Study of Information (IS4SI) are affiliated with *Mathematics* and their members receive a discount on article processing charges.



**16.8 days**

Submission to First Decision Time



**2.592**

Impact Factor in 2021



**High Visibility**

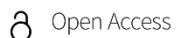
Scopus, SCIE (Web of Science)



**Journal Rank**

JCR—Q1 (*Mathematics*) /  
CiteScore—Q1 (*General Mathematics*)

## Author Benefits



Open Access

Unlimited and free access for readers



No Copyright Constraints

Retain copyright of your work and free use of your article



Thorough Peer-Review



Coverage by Leading Indexing Services



SCIE—Science Citation Index Expanded (Web of Science) and Scopus



No Space Constraints, No Extra Space or Color Charges

No restriction on the maximum length of the papers



Discounts on Article Processing Charges (APC)

If you belong to an institute that participates with the MDPI

Institutional Open Access Program



MDPI – Academic Open Access Publishing since 1996

Switzerland · China · Spain · Romania · Serbia · UK · Japan · Poland · Canada

[www.mdpi.com](http://www.mdpi.com)

## **Springer Book on Big Data Analysis, Biostatistics and Bioinformatics**

Dear Researchers,

Thank you very much for participating in ICSA 2023 China Conference and sharing with us your cutting-edge research. We hope you enjoyed the successful meeting held in the beautiful city of Chengdu, China.

Professor Din Chen, the editor of Springer/ICSA Book Series in Statistics, would like to showcase the scientific output for the conference by making a book, which reflects new challenges and advances in Big Data Analysis, Biostatistics and Bioinformatics. We welcome submissions from all areas of statistics, data science and interdisciplinary areas. Submitted papers are expected to present new methods in statistics and data science, new theories in big data analysis, biostatistics, and applications in bioinformatics.

Below are our plans in details:

1. You are encouraged to submit your research that is related to big data analysis, biostatistics and bioinformatics.
2. To include high quality papers to the book, all submissions will be subject to peer review. Each submission will also be independently reviewed by the reviewers and co-editors. The final accepted papers will be those selected by the co-editors.
3. To make the process move very smoothly,
  - a. If you wish to submit a paper, we need you to indicate your intent of submission by **September 15, 2023**, with a tentative title of the paper (you can change the title later), your name (first,

middle and last), and your affiliation. Please email this information to Yichuan Zhao at [yichuan@gsu.edu](mailto:yichuan@gsu.edu).

b. A manuscript is sent to Dr. Yichuan Zhao by **December 31, 2023**. The decision will be reached by **July 31, 2024**.

c. The book is expected to appear on **November 30, 2024**.

We look forward to hearing from you.

Sincerely yours,

Co-Editors of the book:

Yichuan Zhao, Georgia State University, Atlanta

Din Chen, Arizona State University, Phoenix

## ***Journal of Nonparametric Statistics* Special Issue Submission Information**

Our special issue title for 2023 ICSA China Conference has been added to the author submission form. Authors will now be able to choose the special issue title “2023 ICSA China Conference” from a dropdown menu when submitting to the journal via Submission Portal. The title of the special issue is “**Nonparametric Methods in Data Science and Applications**”.

***Journal of Nonparametric Statistics*** provides a medium for the publication of research and survey work in nonparametric statistics and related areas. The scope includes, but is not limited to the following topics:

- Nonparametric modeling
- Nonparametric function estimation
- Rank and other robust and distribution-free procedures
- Resampling methods
- Lack-of-fit testing
- Multivariate analysis
- Inference with high-dimensional data
- Dimension reduction and variable selection
- Methods for errors in variables, missing, censored, and other incomplete data structures
- Inference of stochastic processes
- Sample surveys
- Time series analysis
- Longitudinal and functional data analysis
- Nonparametric Bayes methods and decision procedures
- Semiparametric models and procedures
- Statistical methods for imaging and tomography
- Statistical inverse problems
- Financial statistics and econometrics
- Bioinformatics and comparative genomics
- Statistical algorithms and machine learning.

Both the theory and applications of nonparametric statistics are covered in the journal. Research applying nonparametric methods to medicine, engineering, technology, science and humanities is welcomed, provided the novelty and quality level are of the highest order.

Authors are encouraged to submit supplementary technical arguments, computer code, data analysed in the paper or any additional information for online publication along with the published paper.

## Options for traveling to Longemont Hotel, Chengdu

**Fly to Chengdu Tianfu International Airport.** Chengdu Tianfu International Airport is the largest civilian airport put into operation in the first year of the "14th Five-Year Plan". It is positioned as the main international aviation hub airport in Chengdu. You can take Bus Rapid Transit Line 1 and get off at Chengdu East Railway Station. From there, it's about a 500-meter walk to the Chengdu Longemont Hotel. Alternatively, you can take the subway Line 18 from Terminal 1 or 2 of Tianfu International Airport. After 11 stops, get off at South Railway Station and transfer to the outer loop of Line 7. After six stops, get off at the east exit of Chengdu East Railway Station and walk about 700 meters to the Chengdu Longemont Hotel.

**Fly to Chengdu Shuangliu International Airport.** Chengdu Shuangliu International Airport is located in the southwest of Chengdu City, Sichuan Province, with a distance of about 16 kilometers from the city center. It has a convenient flight network to Europe, America, Africa, Asia, and Oceania. The airport achieves seamless transfer by integrating different transportation modes, including road, rail, and high-speed rail. There are three city terminals outside the airport, which can realize seamless intermodal transportation. You can take the subway Line 10 from Terminal 2 of Chengdu Shuangliu International Airport. After 5 stops, get off at Taipingyuan Station and transfer to the outer loop of Line 7. After 9 stops, get off at the east exit of Chengdu East Railway Station and walk about 700 meters to the Chengdu Longemont Hotel.

**Take trains.** Chengdu has four passenger railway stations, namely Chengdu Station, Chengdu East Station, Chengdu South Station, and Chengdu West Station.

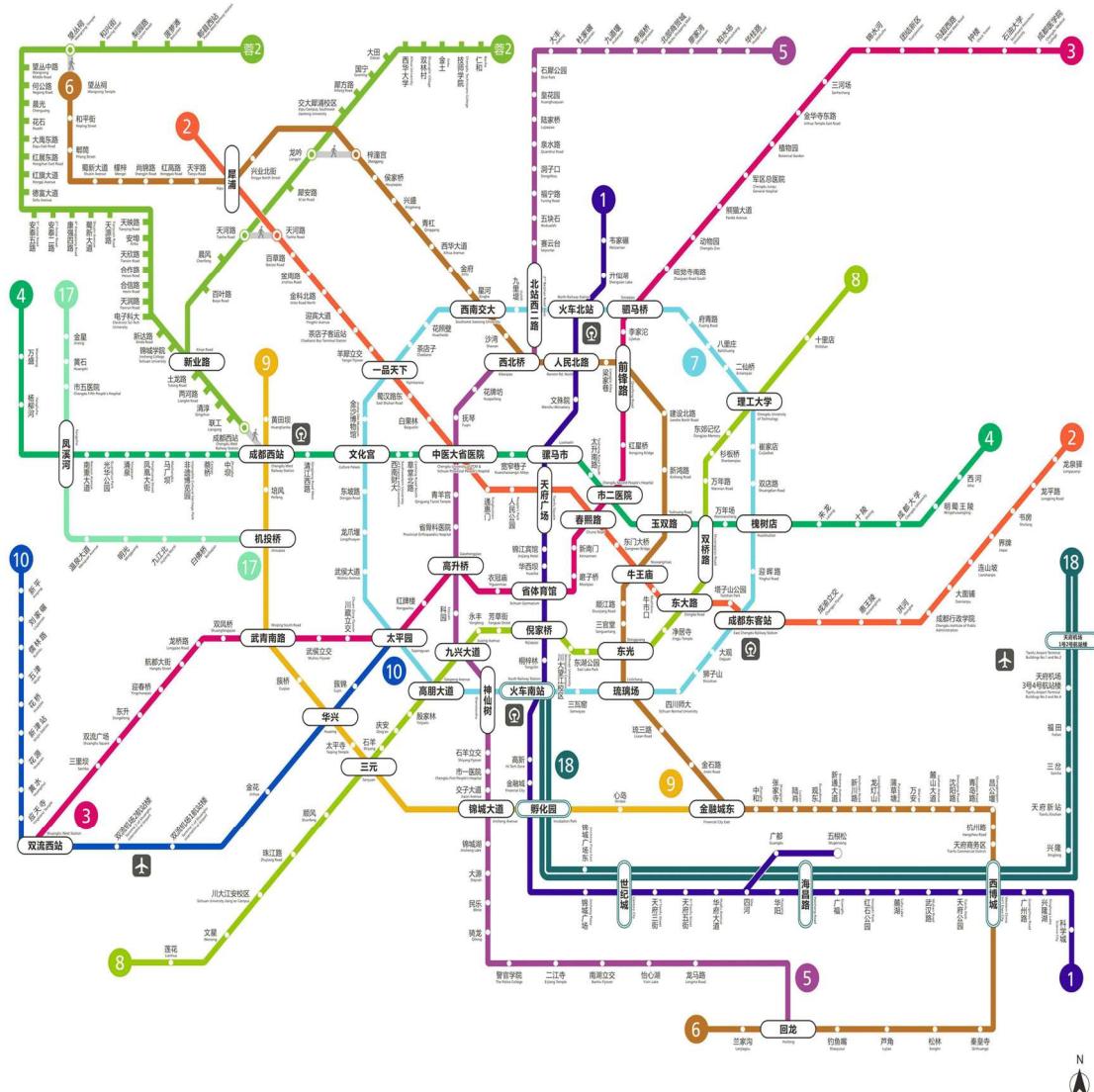
**Chengdu East Station** is one of the largest railway passenger stations in the central and western regions of China. It's only about 400 meters away from Chengdu Longemont Hotel, which takes about 10 minutes to walk there.

**Chengdu South Station** is one of the "three auxiliary stations" in the "three main, three auxiliary" station layout of Chengdu's railway hub plan. It's about 12 kilometers away from Chengdu Longemont Hotel, which takes about 20 minutes by taxi (about 20 RMB).

**Chengdu West Station** is the starting station of the Sichuan-Tibet Railway and is also one of the "three auxiliary stations" in the "three main, three auxiliary" station layout of Chengdu's railway hub plan. It's about 30 kilometers away from Chengdu Longemont Hotel, which takes about 45-60 minutes by taxi (about 70 RMB)



## 成都轨道交通线网图 Chengdu Rail Transit Network Map

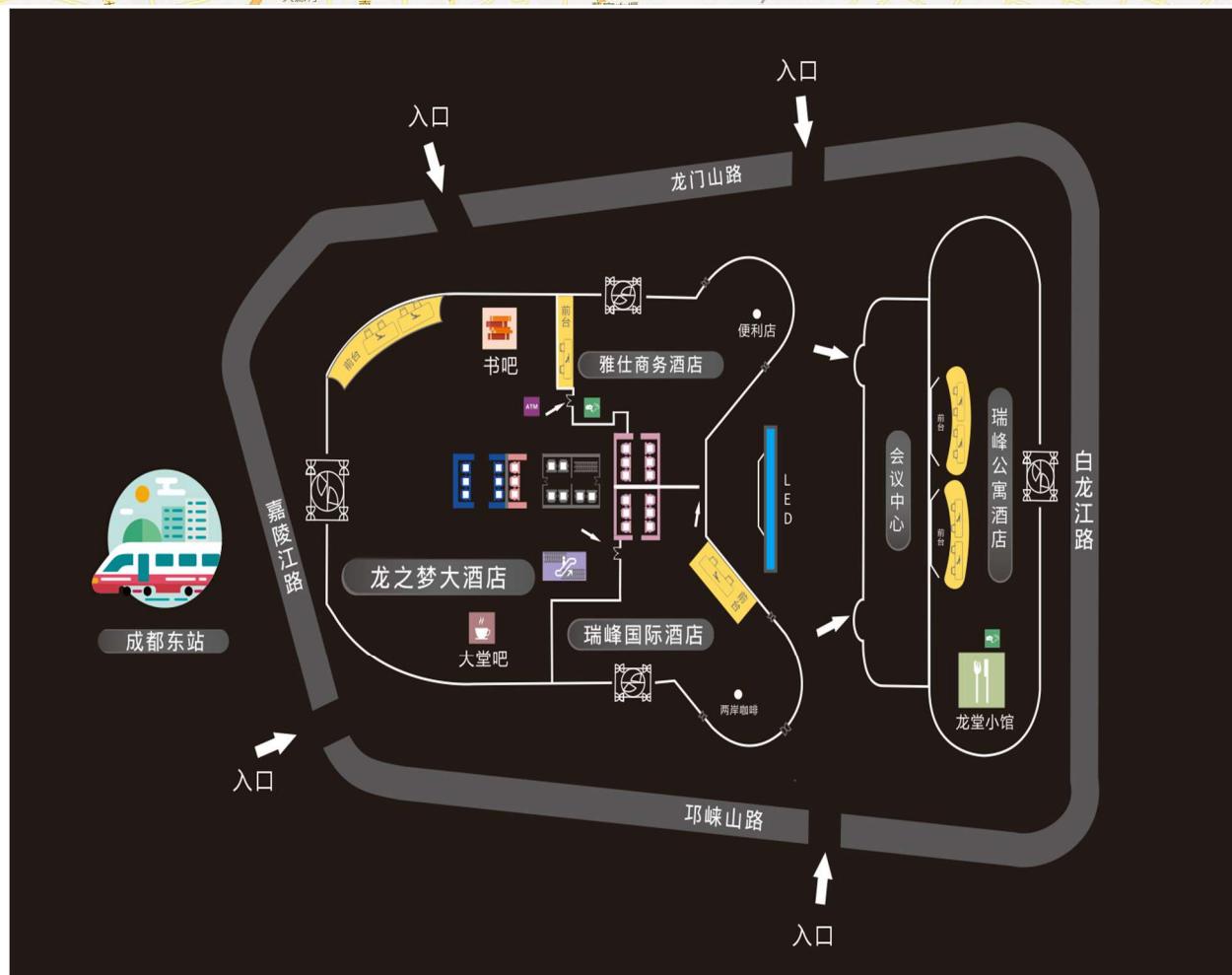
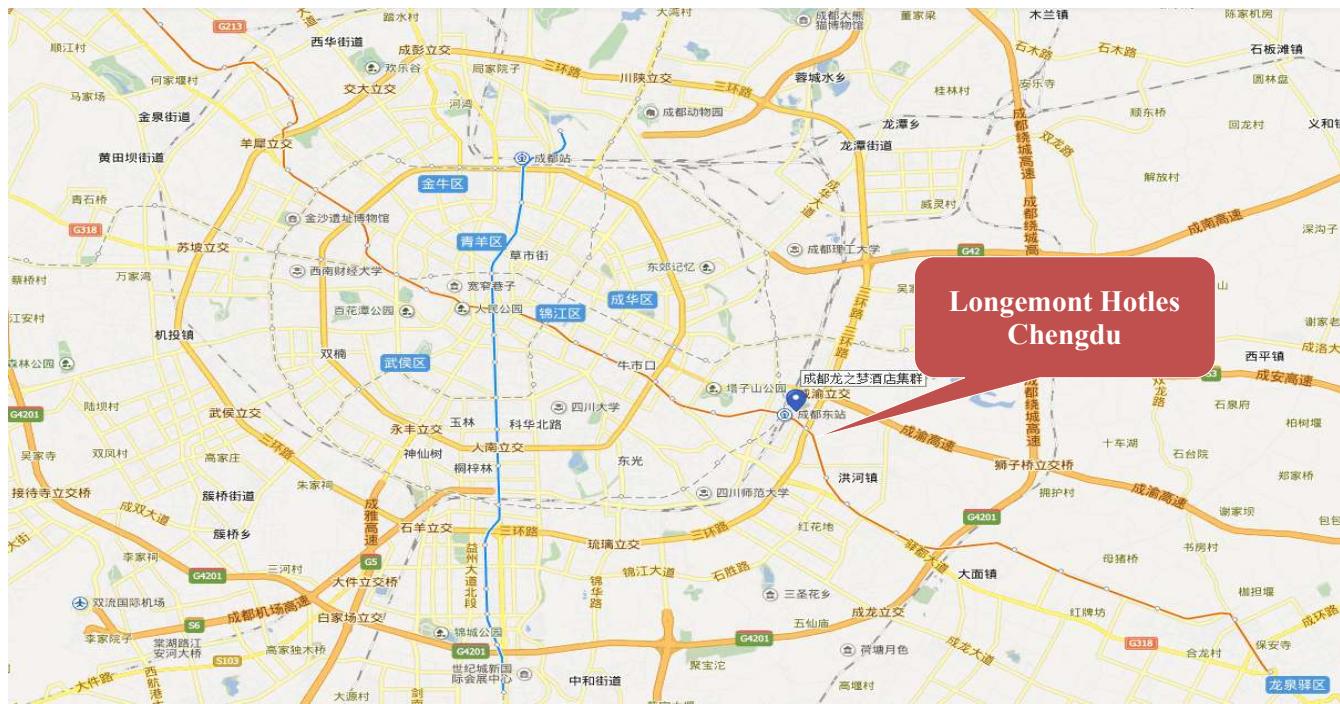


- |                           |                            |   |                        |                        |                        |                        |                        |                        |                           |
|---------------------------|----------------------------|---|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|---------------------------|
| <b>1</b> 1号线<br>Line 1    | <b>2</b> 2号线<br>Line 2     | <b>3</b> 3号线<br>Line 3                        | <b>4</b> 4号线<br>Line 4 | <b>5</b> 5号线<br>Line 5 | <b>6</b> 6号线<br>Line 6 | <b>7</b> 7号线<br>Line 7 | <b>8</b> 8号线<br>Line 8 | <b>9</b> 9号线<br>Line 9 | <b>10</b> 10号线<br>Line 10 |
| <b>17</b> 17号线<br>Line 17 | <b>18</b> 18号快线<br>Line 18 | <b>蓉2</b> 有轨电车<br>蓉2号线<br>Chengdu Tram Line 2 |                        |                        |                        |                        |                        |                        |                           |
| 飞机场<br>The Airport        | 火车站<br>Railway Station     |   |                        |                        |                        |                        |                        |                        |                           |

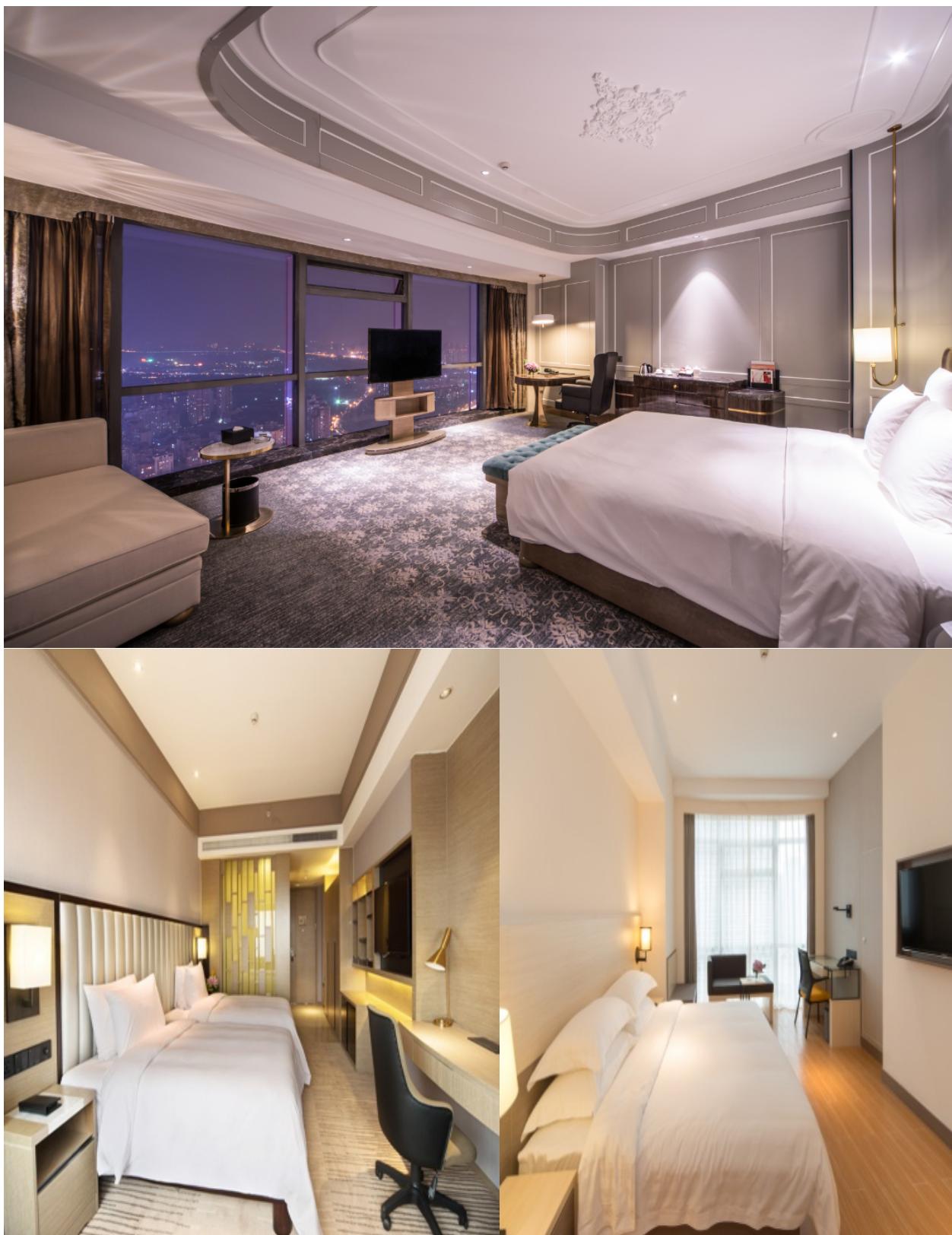
注：此图为变形的折线图，非实际走向

Map of Chengdu Rail

# Transportation

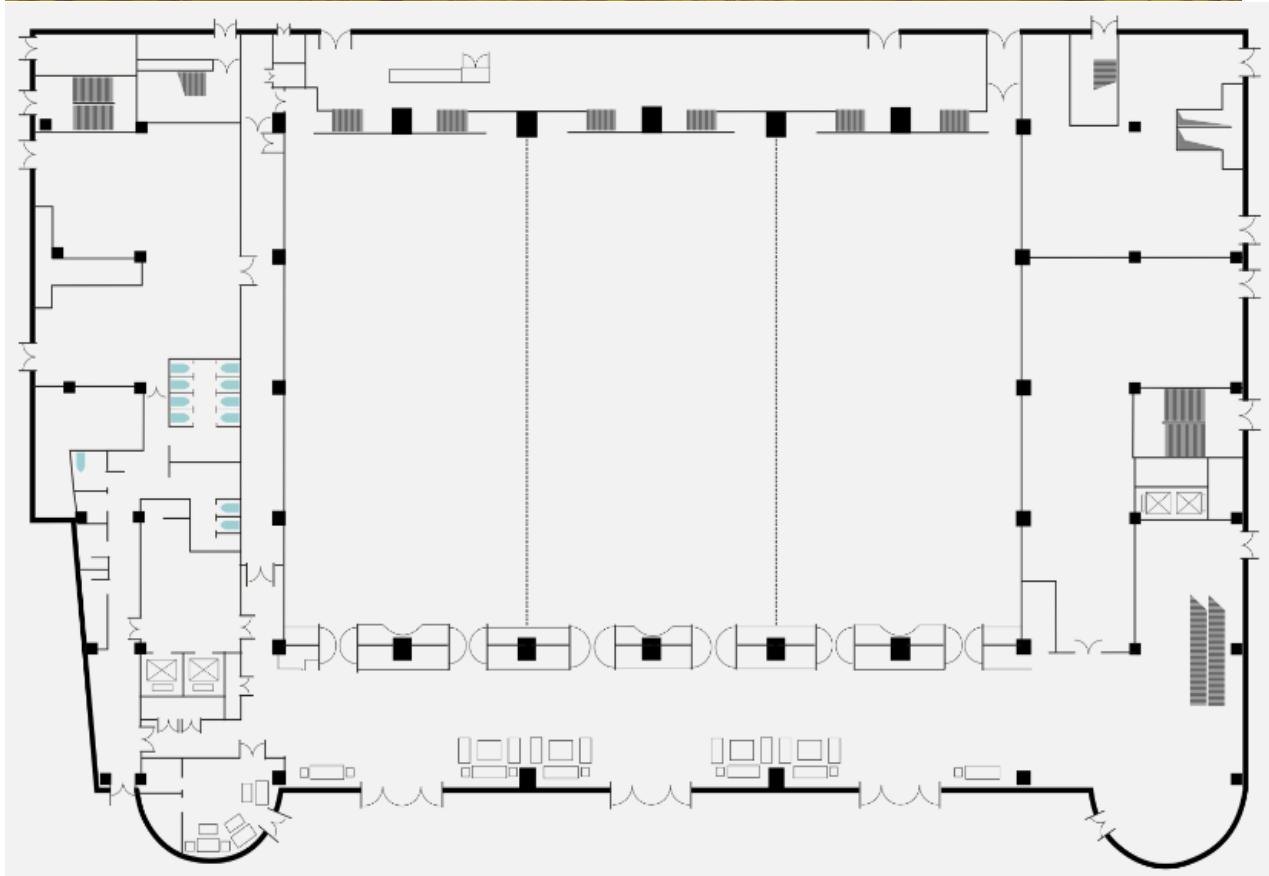


**Location of Longemont Hotel**



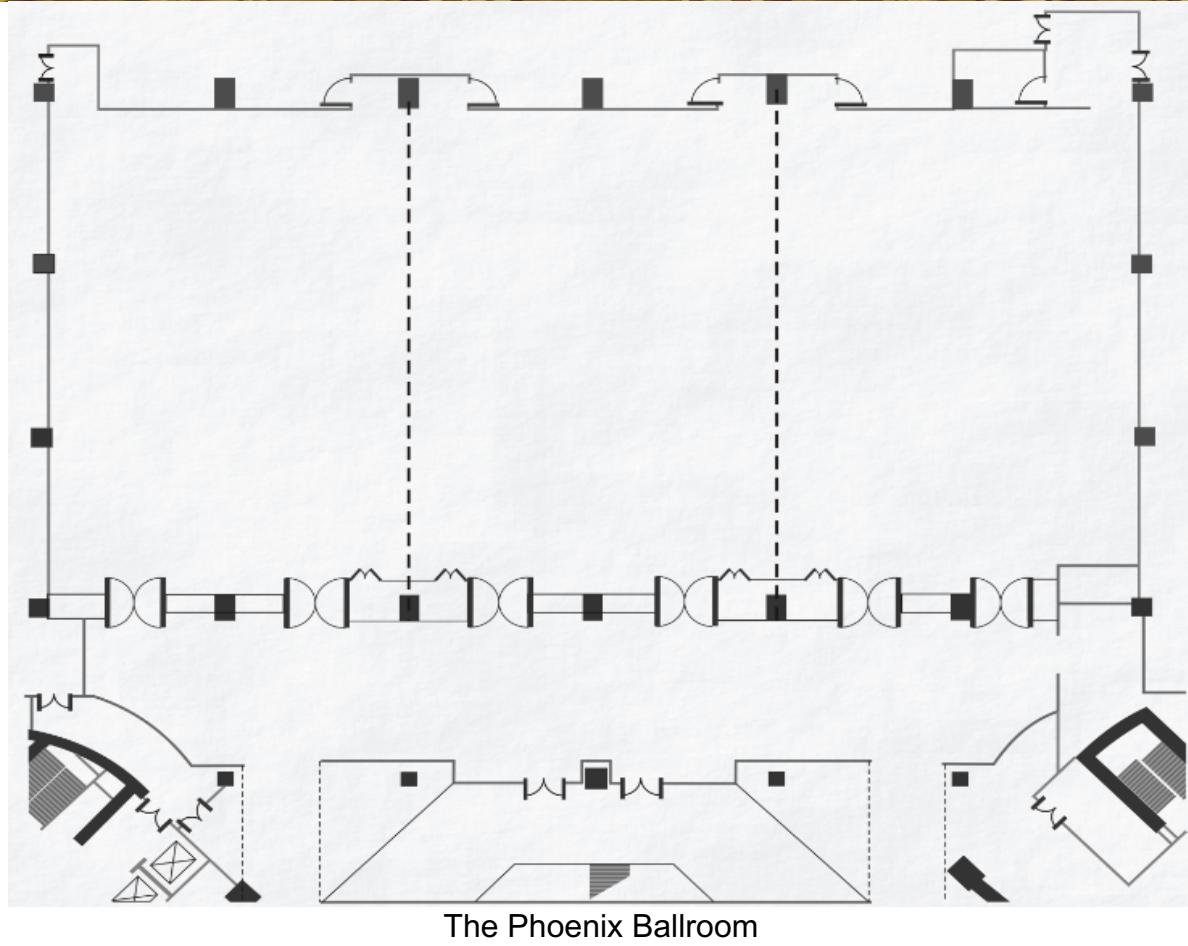
Environment of accommodation  
WIFI: Longemont (Free, without password)

# Conference Venue

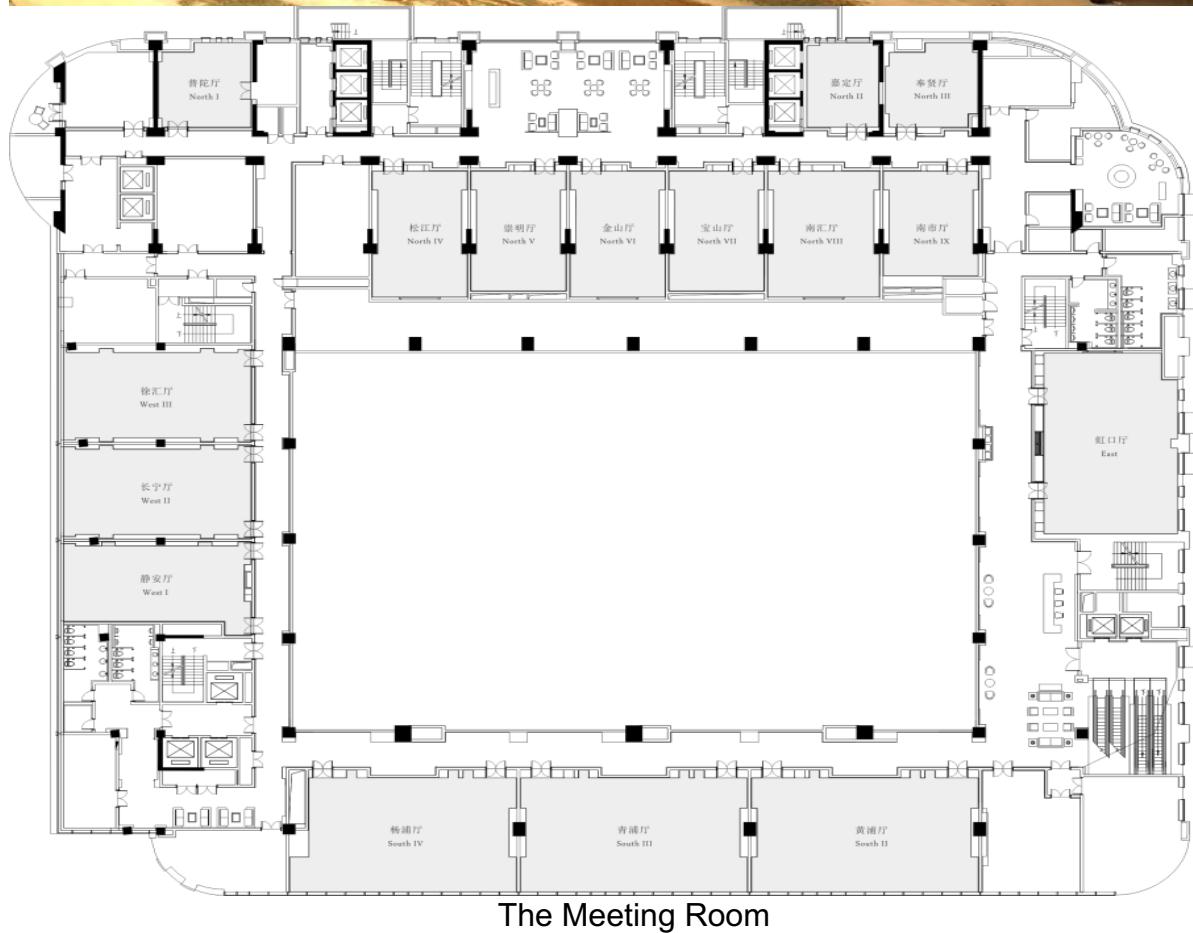


The Dragon Ballroom

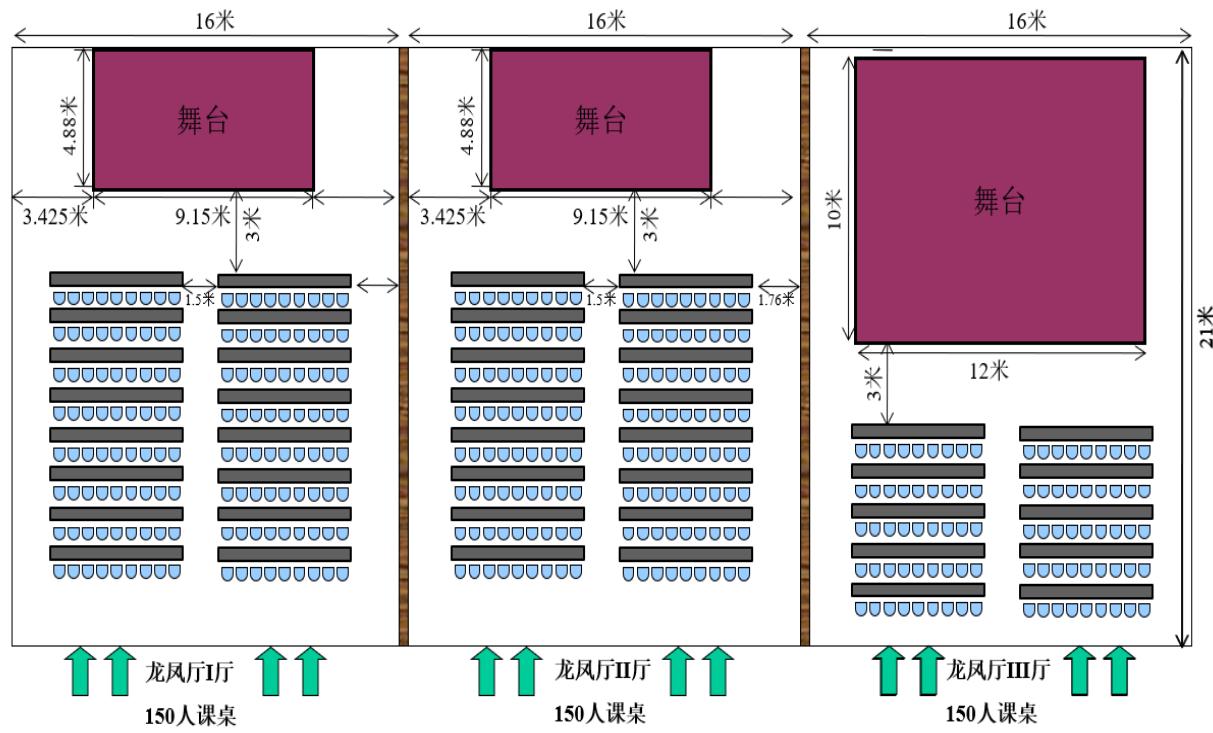
# Conference Venue



# Conference Venue

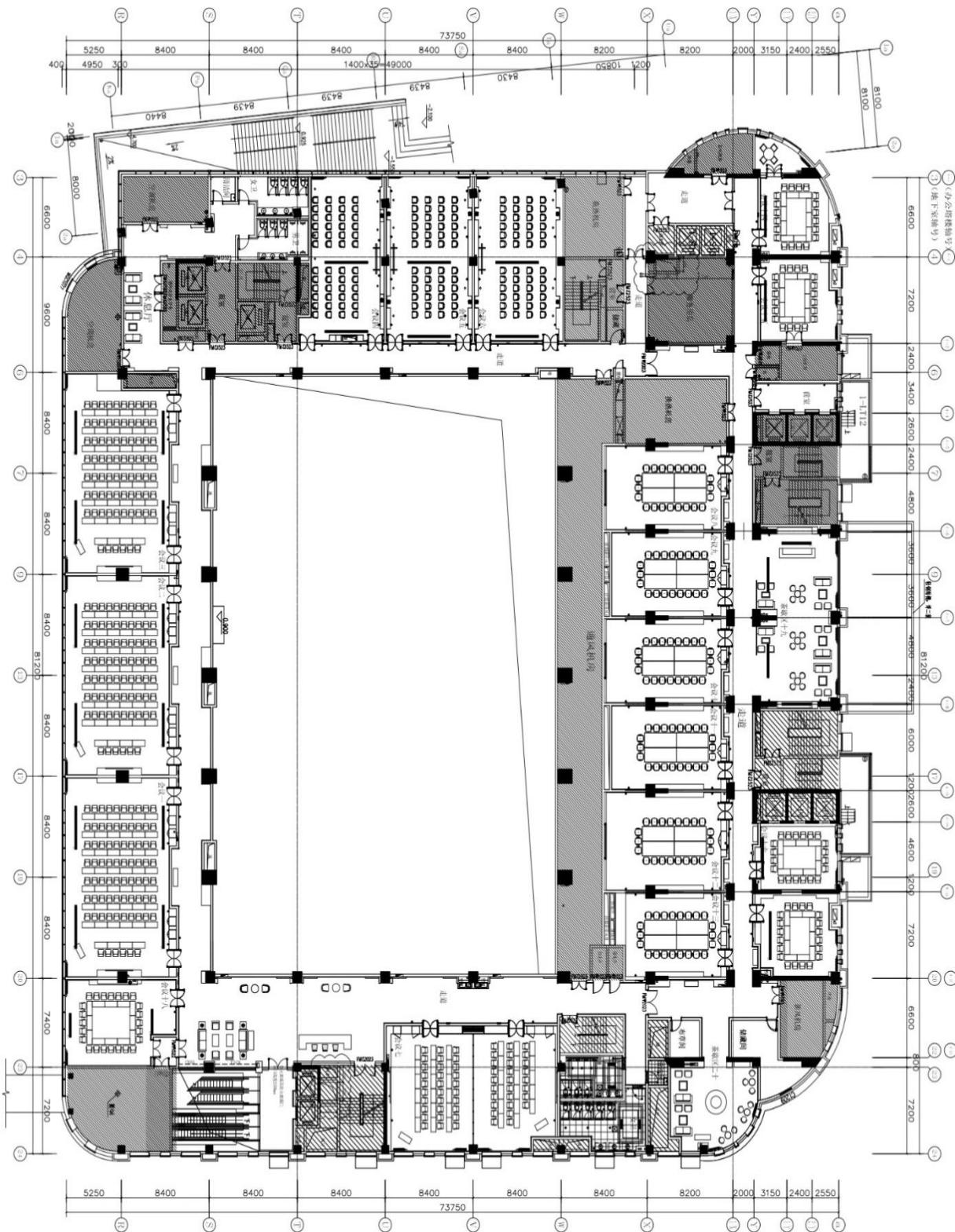


龙凤厅会议平面图



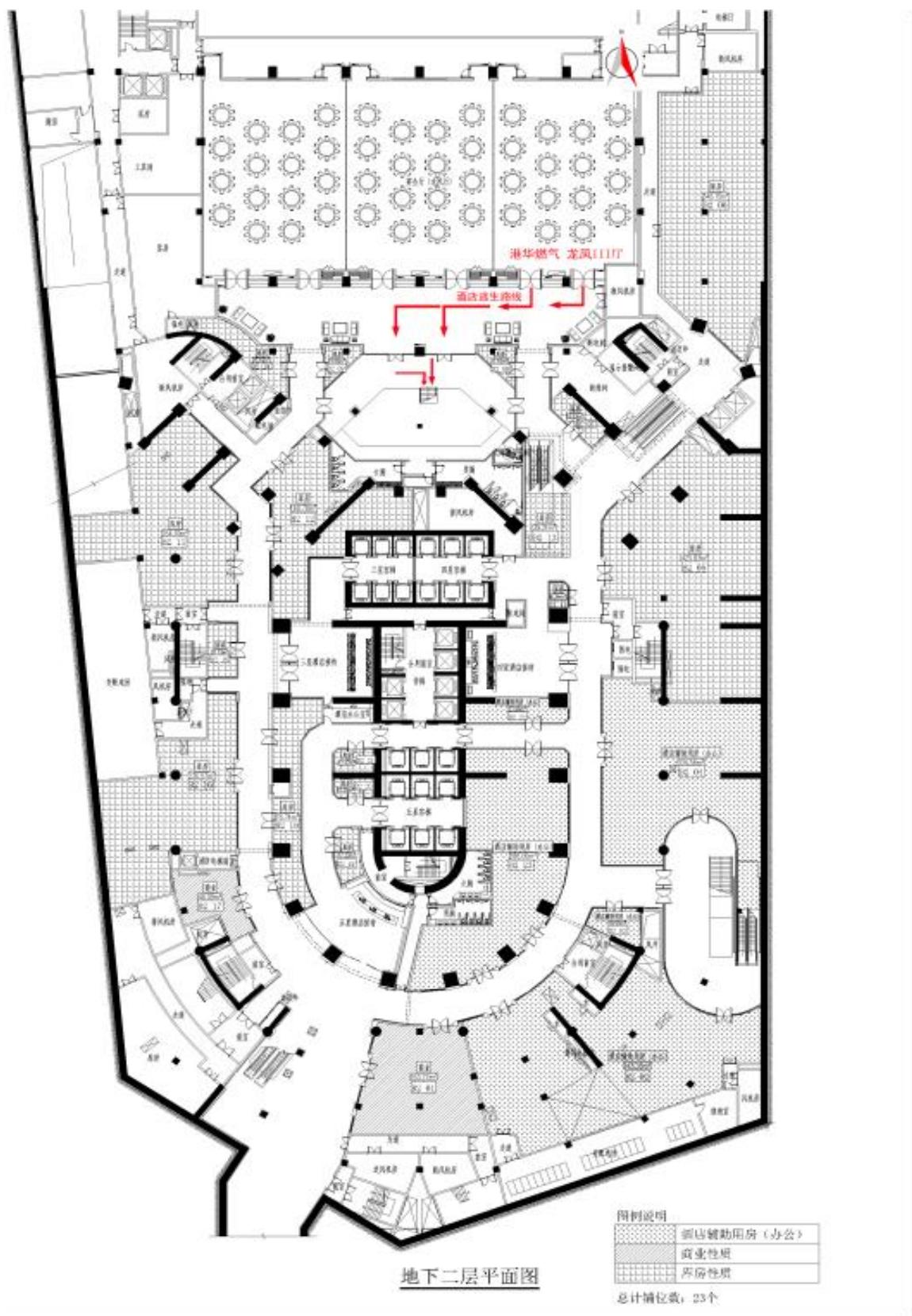
The Longfeng Hall

# Conference Venue



The Second Floor of Longemont Hotel

# Conference Venue



The Second Floor below Ground of Longemont Hotel

## **Opening Mixer:**

The opening mixer will be held on June 30 (Friday), at western dining room of Yashi Business Hotel located on the third floor in Longemont Hotel. The mixer will start at 18:40.

## **Conference Banquet:**

The conference banquet will be held on July 2 (Sunday) night, at Longfeng Hall(龙凤厅) located on the second floor below ground in Longemont Hotel.

The banquet will start at 18:45.

Banquet speaker: **Dr. Zongben Xu**

## **Meals and Tea Breaks:**

In order to facilitate informal discussions among the participants, complimentary meals and tea breaks will be provided.

Two meals (12:00--14:00 for lunch, and 18:40--20:30 for dinner) of July 1 will be served at Yashi Business Hotel and Ruifeng Business Hotel, on the third floor of Longemont Hotel during the conference.

Lunch (12:00--14:00) of July 2 and July 3 will be served at western dining room of Longemont Hotel located on the third floor in Longemont Hotel.

The tea breaks will be held in the lobbies of the conference buildings.

| Name         |       | School |
|--------------|-------|--------|
| Ao           | Li    | SWJTU  |
| Haonan       | Li    | SWJTU  |
| Jinbo        | Wang  | SWJTU  |
| Yu           | Han   | SWJTU  |
| Hang         | Ruan  | SWJTU  |
| Chifeng      | Zhou  | SWJTU  |
| Ziming       | Wang  | SWJTU  |
| Ziyue        | Li    | SWJTU  |
| Lei          | Wang  | SWJTU  |
| Qing         | Li    | SWJTU  |
| Rui          | Wu    | SWJTU  |
| Yufan        | Shi   | SWJTU  |
| Chenchen     | Zhao  | SWJTU  |
| Tongbo       | Yang  | SWJTU  |
| Zhibin       | Mao   | SWJTU  |
| Rui          | Yin   | SWJTU  |
| Shiyu        | Guo   | SWJTU  |
| Jinghui      | Wang  | SWJTU  |
| Yi           | Zhang | SWJTU  |
| Yunlin       | Wu    | SWJTU  |
| Yijun        | Pan   | SWJTU  |
| Shuangshuang | Wu    | SWJTU  |
| Yi           | Wu    | SWJTU  |
| Xinling      | Liu   | SWJTU  |
| Xing         | Wang  | SWJTU  |
| Xin          | Wei   | SWJTU  |
| Yunlong      | Zhao  | SWJTU  |
| Yilong       | CAI   | SWJTU  |
| Huiyuan      | Du    | SWJTU  |
| Yang         | Wang  | SWJTU  |
| Xianhua      | Li    | SWJTU  |
| Xutao        | Li    | SWJTU  |
| Bo           | Wei   | SWJTU  |
| Danting      | Wang  | SWJTU  |
| Yimei        | Zheng | SWJTU  |
| Xuefeng      | CAI   | SWJTU  |
| Xinya        | Li    | SWJTU  |
| Zhi          | Ye    | SWJTU  |
| Yilin        | Chen  | SWJTU  |
| Qingbiao     | Song  | SWJTU  |
| Ying         | Li    | SWJTU  |
| Xiaolong     | Guo   | SWJTU  |
| Wenxue       | Li    | SWJTU  |
| Yijie        | Zhou  | SWJTU  |
| Shan         | Li    | SWJTU  |
| Weizheng     | Gan   | SWJTU  |
| Yue          | Guo   | SWJTU  |
| Wanmei       | Cui   | SWJTU  |
| Hanxun       | Li    | SWJTU  |
| Ruihan       | He    | SWJTU  |
| Yu           | Xia   | SWJTU  |
| Wanli        | Ma    | SWJTU  |
| Fengyi       | Deng  | SWJTU  |
| Heyi         | Yang  | SWJTU  |
| Maojie       | Ye    | SWJTU  |

# Program Overview

| Day 1           |  |                              |                |
|-----------------|--|------------------------------|----------------|
|                 | In person front desk open from 7:30am-7:30pm |                              | Host           |
| 8:00AM-8:30AM   | Virtual Main Room/ local auditorium          | Welcome and Opening Ceremony | Yichuan Zhao   |
| 8:30AM-9:30AM   |  | Keynote Lecture 1            |                |
| 9:30AM-10:00AM  | Coffee Break                                 |                              |                |
| 10:00AM-11:40AM | Breakout rooms/ local session rooms          | Parallel Invited Sessions    | Session Chairs |
| 12:00PM-1:00PM  | Lunch  |                              |                |
| 1:00PM-2:40PM   | Breakout rooms/ local session rooms          | Parallel Invited Sessions    | Session Chairs |
| 2:40PM-3:00PM   | Coffee Break                                 |                              |                |
| 3:00PM-4:40PM   | Breakout rooms/ local session rooms          | Parallel Invited Sessions    | Session Chairs |
| 4:40PM-5:00PM   | Coffee Break                                 |                              |                |
| 5:00PM-6:40PM   | Breakout rooms/ local session rooms          | Parallel Invited Sessions    | Session Chairs |
| 4:00PM-7:40PM   | Poster Session                               |                              |                |
| Day 2           |  |                              |                |
|                 | In person front desk open from 7:30am-7:30pm |                              | Host           |
| 8:30AM-9:30AM   | Virtual Main Room/ local auditorium          | Keynote Lecture 2            | Zhezhen Jin    |
| 9:30AM-10:00AM  | Coffee Break                                 |                              |                |
| 10:00AM-11:40AM | Breakout rooms/ local session rooms          | Parallel Invited Sessions    | Session Chairs |
| 12:00PM-1:00PM  | Lunch  |                              |                |
| 1:00PM-2:40PM   | Breakout rooms/ local session rooms          | Parallel Invited Sessions    | Session Chairs |
| 2:40PM-3:00PM   | Coffee Break                                 |                              |                |
| 3:00PM-4:40PM   | Breakout rooms/ local session rooms          | Parallel Invited Sessions    | Session Chairs |
| 4:40PM-5:00PM   | Coffee Break                                 |                              |                |
| 5:00PM-6:40PM   | Breakout rooms/ local session rooms          | Parallel Invited Sessions    | Session Chairs |
| 6:45PM-8:45PM   | Conference Banquet                           |                              |                |
| Day 3           |  |                              |                |
|                 | In person front desk open from 8:30am-1:00pm |                              | Host           |
| 8:30AM-10:10AM  | Breakout rooms/ local session rooms          | Parallel Invited Sessions    | Session Chairs |
| 10:10AM-10:30AM | Coffee Break                                 |                              |                |
| 10:30AM-12:10PM | Breakout rooms/ local session rooms          | Parallel Invited Sessions    | Session Chairs |
| 12:00PM-1:00PM  | Lunch  |                              |                |

# Keynote Lectures



**Dr. Jun Liu** is a Professor of Statistics at Harvard University, with a courtesy appointment at Harvard School of Public Health. Dr. Liu received his BS degree in mathematics in 1985 from Peking University and Ph.D. in statistics in 1991 from the University of Chicago. He held Assistant, Associate, and full professor positions at Stanford University from 1994 to 2003. Dr. Liu received the NSF CAREER Award in 1995 and the Mitchell Award in 2000. In 2002, he won the prestigious COPSS Presidents' Award (given annually to one individual under age 40). He was selected as a Medallion Lecturer in 2002, a Bernoulli Lecturer in 2004, a Kuwait Lecturer of Cambridge University in 2008; and elected to Fellow of the Institute of Mathematical Statistics in 2004, Fellow of the American Statistical Association in 2005, and Fellow of the

International Society for Computational Biology in 2022. He was awarded the Morningside Gold Medal in Applied Mathematics in 2010 (once every 3 years to an individual of Chinese descent under age 45). He was honored with the Outstanding Achievement Award and the Pao-Lu Hsu Award (once every 3 years) by the International Chinese Statistical Association in 2012 and 2016, respectively. In 2017, he was recognized by the Jerome Sacks Award for outstanding Cross-Disciplinary Research.

**Time and Location:** 8:30am-9:30am, July 1 (Beijing Time), LONGFENG HALL

**Host:** Yichuan Zhao

**Title:** On Methods for Controlled Variable Selection in linear, generalized-linear, and index models

**Abstract:** A classical statistical idea is to introduce data perturbations and examine their impacts on a statistical procedure. In the same token, the knockoff methods carefully create “matching” fake variables in order to measure how real signals stand out. I will discuss some recent investigations we made regarding both methodology and theory on a few related methods applicable to a wide class of regression models including the knock off filter, data splitting (DS), Gaussian mirror (GM), for controlling false discovery rate (FDR) in fitting linear, generalized linear and index models. We theoretically compare, under the weak-and-rare signal framework for linear models, how these methods compare with the oracle OLS method. We then focus on the DS procedure and its variation, Multiple Data Splitting (MDS), which is useful for stabilizing the selection result and boosting the power. DS and MDS are straightforward conceptually, easy to implement algorithmically, and applicable to a wide class of linear and nonlinear models. Interestingly, their specializations in GLMs result in scale-free procedures that can circumvent difficulties caused by non-traditional asymptotic behaviors of MLEs in moderate-dimensions and debiased Lasso estimates in high-dimensions. For index models, we had developed an earlier LassoSIR algorithm (Lin, Zhao and Liu 2019), which fits the DS framework quite well. I will also discuss some applications and open questions. The presentation is based on joint work with Chenguang Dai, Buyu Lin, Xin Xing, Tracy Ke, Yucong Ma, and Zhigen Zhao.

# Keynote Lectures



**Dr. Fang Yao** is Chair Professor in School of Mathematical Sciences, Director of Center for Statistical Science at Peking University. He is a Fellow of IMS and ASA, and an elected member of ISI. He received his B.S. degree in 2000 from University of Science & Technology in China, and his Ph.D. degree in Statistics in 2003 at UC Davis. He was a tenured Full Professor in Statistical Sciences at University of Toronto.

Dr. Yao's research primarily focuses on functional and longitudinal data, complex structures such as high dimensions, manifolds and dynamics, and their applications in various disciplines. In 2014, he received the CRM-SSC Prize that recognizes a statistical scientist's professional accomplishments in research primarily conducted in Canada. He has served as the Editor for Canadian Journal of Statistic, and is/was on editorial boards for nine statistical journals, including AOS and JASA.

**Time and Location:** 8:30am-9:30am, July 2 (Beijing Time), LONGFENG HALL

**Host:** Zhezhen Jin

**Title:** Theory of FPCA for discretized functional data

**Abstract:** Functional data analysis is an important research field in statistics which treats data as random functions drawn from some infinite-dimensional functional space, and functional principal component analysis (FPCA) plays a central role for data reduction and representation. After nearly three decades of research, there remains a key problem unsolved, namely, the perturbation analysis of covariance operator for diverging number of eigencomponents obtained from noisy and discretely observed data. This is fundamental for studying models and methods based on FPCA, while there has not been much progress since the result obtained by Hall et al. (2006) for a fixed number of eigenfunction estimates. In this work, we establish a unified theory for this problem, deriving the moment bounds of eigenfunctions and asymptotic distributions of eigenvalues for a wide range of sampling schemes. We also exploit double truncation to derive the uniform convergence of such estimated eigenfunctions. The technical arguments in this work are useful for handling the perturbation series of discretely observed functional data and can be applied in models and methods involving inverse using FPCA as regularization, such as functional linear regression.



**Zongben Xu** is a professor in mathematics and computer science at Xi'an Jiaotong University. He received his Ph.D. degrees in mathematics from Xi'an Jiaotong University, China, in 1987. His current research interests include applied mathematics and mathematical methods of big data and artificial intelligence. He established the L(1/2) regularization theory for sparse information processing. He also found and verified Xu-Roach Theorem in machine learning, and established the visual cognition based data modelling principle, which have been widely applied in scientific and engineering fields. He initiated several mathematical theories, including the non-logarithmic transform based CT model, and ultrafast MRI imaging, which provide principles and technologies for the development of a new generation of intelligent medical imaging equipment. He is owner of the Hua Loo-keng Prize of Mathematics in 2022, Tan Kan Kee Science Award in Science Technology in 2018, the National Natural Science Award of China in 2007, and winner of CSIAM Su Buchin Applied Mathematics Prize in 2008. He delivered a 45-minute talk on the International Congress of Mathematicians 2010. He was elected as member of Chinese Academy of Science in 2011.

Zongben Xu was the vice-president of Xi'an Jiaotong University. He currently makes several important services for government and professional societies, including the director for Pazhou Lab (Huangpu), director for the National Engineering Laboratory for Big Data Analytics, a member of National Big Data Expert Advisory Committee and the Strategic Advisory Committee member of National Open Innovation Platform for New Generation of Artificial Intelligence.

**Time and Location:** 7:30 pm-8:15 pm, July 2 (Beijing Time), LONGFENG HALL

**Host:** Yichuan Zhao

**Title:** Viewing Statistics from a Data Science Perspective

**Abstract:** In the era of big data, characterized by digitization, networking, and intelligence, data serves as both a production factor and a powerful tool for scientific discovery. Statistics has long been considered as the scientific foundation and methodology, guiding and leading the generation, analysis, and utilization of data. Can this guiding and leading role continue in the age of big data? The statistics community is supposed to ponder and answer this question. This talk will discuss this issue.

Big data nurtures data science, and data science carries the future of big data. First, we rigorously define data science, elaborating on its multidisciplinary attributes, advanced nature, "three transformations" connotation, and unique disciplinary methodology of "modeling, analysis, computation, and learning fusion." Secondly, we compare the

## Banquet Talk

similarities and differences between data science and statistics, pointing out the unique contributions of statistics to data science and its "core" role in the development of data science. Based on this, we outline the limitations of statistics in terms of research objects, methods, and values. We provide examples to demonstrate that the integration of data science and statistics can inspire new research questions and methods in various fields, thereby promoting the novel development of data science. Consequently, we assert that statistics can continue to guide and lead data science disciplines in the big data era if it proactively embraces and strides towards data science.

# Junior Researcher Awards

## ICSA China Conference Junior Researcher Awards:

### **Erdmann-Pham Dan, Stanford University**

- Title: Exact and Efficient Multivariate Two-Sample Tests through Generalized Linear Rank Statistics
- Time: July 3, 10:30-12:10
- Session 23CHI172: Junior Researcher Award Session
- Room: Jiading Hall

### **Feiyu Jiang, Fudan University**

- Title: Testing Serial Independence of Object-Valued Time Series
- Time: July 3, 10:30-12:10
- Session 23CHI172: Junior Researcher Award Session
- Room: Jiading Hall

### **Cheng Meng, Renmin University of China**

- Title: Importance Sparsification for Sinkhorn Algorithm
- Time: July 3, 10:30-12:10
- Session 23CHI172: Junior Researcher Award Session
- Room: Jiading Hall

### **Yuting Wei, University of Pennsylvania**

- Title: The Lasso with General Gaussian Designs with Applications to Hypothesis Testing
- Time: July 1, 13:00-14:40
- Session 23CHI134: Message Passing and Differential Privacy
- Room: Xuhui Hall

### **Wenzhuo Zhou, University of California Irvine**

- Title: Distributional Shift-Aware Off-Policy Interval Estimation: A Unified Error Quantification Framework
- Time: July 3, 10:30-12:10
- Session 23CHI172: Junior Researcher Award Session
- Room: Jiading Hall

# Poster Session

| No | First Name | Last Name | Title   |
|----|------------|-----------|---|
| 1  | Zhexiao    | Lin       | On regression-adjusted imputation estimators of the average treatment effect  |
| 2  | Qingwen    | Zhang     | Sobolev Calibration of Imperfect Computer Models  |
| 3  | Jiahang    | JIANG     | Adaptive Learning with Random Smoothing Kernel Hilbert Spaces   |
| 4  | Jiuzhou    | Wang      | Multiple Augmented Reduced Rank Regression for Pan-Cancer Analysis  |
| 5  | Wei        | Liang     | Propensity Score Weighting with Post-Treatment Survey Data  |
| 6  | IIRONG     | ZHU       | Constructing genotype and phenotype network helps reveal disease heritability and genome-wide association studies                       |
| 7  | Yu         | Shi       | Mixed latent graphical models with mixed measurement error and misclassification in variables   |
| 8  | Rong       | Li        | Regulation-incorporated Gene Expression Network-based Heterogeneity Analysis  |
| 9  | Yicheng    | Zeng      | Sketched Ridgeless Linear Regression: The Statistical Role of Sketching   |
| 10 | JIYUAN     | HU        | Accurate estimation of breakpoints in piecewise linear mixed-effects models with application to longitudinal ophthalmic studies         |
| 11 | Iris       | Sun       | Immuno-bridging in vaccine development  |
| 12 | Sifan      | Liu       | New Perspectives of p-Value and its Strength of Evidence Measured by Confidence Distribution  |
| 13 | Yiran      | Hu        | Estimand implementation in the time related endpoints and analysis in clinical trials   |
| 14 | Quan       | Sun       | Improving Polygenic Risk Prediction in Admixed Populations by Explicitly Modeling Ancestral-Differential Effects via GAUDI              |
| 15 | Zhiqiang   | Liao      | Overfitting reduction in convex regression  |
| 16 | Zixuan     | Xu        | Conditional Independence Test Based on Conditional Density-Ratio Estimation   |
| 17 | Yupeng     | Zhang     | Modelling Misclassification Errors in Scores Data   |
| 18 | Haoze      | Hou       | Non-parametric Estimation and Inference of General Heterogeneous Causal Effects with Covariate Measurement Error                        |
| 19 | Hongwei    | Shi       | Tests for ultrahigh-dimensional partially linear regression models  |
| 20 | Yulin      | Zhang     | Semiparametric Estimation and Inference for Continuous Treatment Models   |
| 21 | Yuliang    | Shi       | Variable Selection for Causal Modeling in Missing Exposure Problems   |
| 22 | Hongyu     | Lai       | A comparison of methods for handling confounder missingness from graphical model perspectives   |
| 23 | zili       | liu       | Simultaneous Variable Selection and Estimation of Survival Data with Informative Censoring  |
| 24 | Chengmei   | Niu       | Reducing the inversion bias of random sampling  |
| 25 | Yongqi     | Du        | Precise Expression of Neural Tangent Kernel for High-dimensional Gaussian Mixture Data with Application to DNN Training and Compression |
| 26 | Daolin     | Pang      | Factor augmented inverse regression and its application to microbiome data analysis   |
| 27 | Yuxuan     | Zong      | ZINBA: a zero-inflated negative binomial framework for microbiome absolute abundance analysis   |
| 28 | Xiaoxiu    | Tan       | mDriver: identifying driver microbes in the microbial community based on time series data   |

# Poster Session

|    |          |          |   |
|----|----------|----------|---|
| 29 | Xinbo    | Wang     | HILAMA: High-dimensional multi-omic mediation analysis with latent confounding  |
| 30 | Sanyou   | WU       | Individualized Region Detection via Multiple Instance Learning  |
| 31 | Miaomiao | Su       | A Two-Stage Optimal Subsampling Estimation for Missing Data Problems with Large-Scale Data  |
| 32 | Guang    | Yang     | Region Detection and Image Clustering via Sparse Kronecker Product Decomposition  |
| 33 | Ruoyu    | Wang     | Debiased estimating equation method for summary statistics-based Mendelian randomization  |
| 34 | Di       | Wang     | Incorporating External Risk Information with the Cox Model under Population Heterogeneity: Applications to Trans-Ancestry Polygenic Hazard Scores |
| 35 | Lingfeng | Luo      | Coordinate Ascent-Based Gradient Boosting for Identifying Time-varying Effects and Interaction Terms with a Discrete Failure Time Model           |
| 36 | Qishuo   | Yin      | Optimal Sample Splitting for Multiple Outcomes Treatment Effect Detection   |
| 37 | Daniel   | Phillips | A Bayesian joint model for COVID-19 antibody decay and risk of infection  |
| 38 | Zhijun   | Liu      | A CLT for the LSS of large dimensional sample covariance matrices with diverging spikes   |
| 39 | Dazhi    | Zhao     | Novel Empirical Likelihood Methods for the Cumulative Hazard Ratio  |
| 40 | Shushan  | Wu       | Subsampling in Large Graphs Using Ricci Curvature   |
| 41 | Luyang   | Fang     | SPOT: An Active Learning Framework for Deep Neural Networks   |
| 42 | Daiqi    | Gao      | Non-asymptotic Properties of Individualized Treatment Rules from Sequentially Rule-Adaptive Trials  |
| 43 | Haoran   | Lu       | Mortgage Prerepayment Modeling via a Smoothing Spline State Space Model   |
| 44 | Xingyun  | Cao      | Exact confidence interval based on plausibility function for one binomial proportion  |
| 45 | Xirui    | Liu      | Selection of fixed effects in finite mixture of high-dimensional linear mixed-effects models  |
| 46 | Xiangyu  | Shi      | Max-sum test based on Spearman's footrule for high-dimensional independence tests   |
| 47 | Liqi     | Xia      | An improvement to the new coefficient of Chatterjee   |
| 48 | Yangbing | Tang     | Rank-based instrumental variable estimation for semiparametric varying coefficient spatial autoregressive models                                  |
| 49 | Jiazhang | Cai      | WEST: An Ensemble Method for Spatial Transcriptomics Analysis   |

## Scientific Program (June 30 - July 3)

---

### July 1, 8:00-9:30

#### **Session 23CHIKT1: Opening Ceremony and Keynote Lecture 1**

Room: LONGFENG HALL

Organizer: ICSA China Conference Organizing Committee.

Chair: Yichuan Zhao, Georgia State University.

8:00 Welcome and Opening Ceremony

8:30 Methods for Controlled Variable Selection in Linear, Generalized-Linear, and Index Models

*Jun Liu.* Harvard University

### July 1, 10:00-11:40

#### **Session 23CHI100: Challenges in the Analysis of Survival Data with Complex Features**

Room: XUHUI HALL

Organizer: Wenqing He, University of Western Ontario.

Chair: Wenqing He, University of Western Ontario.

10:00 Separable Pathway Effects of Semi-Competing Risks via Multi-State Models

*Yuhao Deng<sup>1</sup>, ♦Yi Wang<sup>2</sup>, Xiang Zhan<sup>2</sup> and Xiao-Hua Zhou<sup>3</sup>.* <sup>1</sup>School of Mathematical Sciences, Peking University <sup>2</sup>Beijing International Center for Mathematical Research, Peking University <sup>3</sup>Beijing International Center for Mathematical Research, and Department of Biostatistics, School of Public Health, Peking University

10:25 Zero-Inflated Poisson Models with Measurement Error in Response

*Grace Yi.* University of Western Ontario

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

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

11:15 Multi-Task Prediction Model for Survival Data

*Shuai You<sup>1</sup>, Xiaowen Cao<sup>1</sup>, Grace Yi<sup>2</sup>, ♦Xuekui Zhang<sup>1</sup> and Li Xing<sup>3</sup>.* <sup>1</sup>University of Victoria <sup>2</sup>Western University <sup>3</sup>University of Saskatchewan

11:40 Floor Discussion.

#### **Session 23CHI115: Recent Advanced of Modeling to Complex Censored Data**

Room: BAOSHAN HALL

Organizer: Chunjie Wang, Changchun University of Technology.

Chair: Chunjie Wang, Changchun University of Technology.

10:00 Regression Analysis for the Semi-Competing Models of Interval-Censored Data with a Cured Subgroup

*Yichen Lou<sup>1</sup>, ♦Peijie Wang<sup>1</sup> and Jianguo Sun<sup>2</sup>.* <sup>1</sup>Jilin University <sup>2</sup>University of Missouri

10:25 Joint Analysis of Informatively Interval-Censored Failure Time and Panel Count Data

<sup>♦</sup>*Shuying Wang<sup>1</sup>, Chunjie Wang<sup>1</sup>, Xinyuan Song<sup>2</sup> and Da Xu<sup>3</sup>.* <sup>1</sup>Changchun University of Technology <sup>2</sup>The Chinese University of Hong Kong <sup>3</sup>Northeast Normal University

10:50 Smoothed Estimation on Optimal Treatment Regime in Semi-Supervised Setting

*Xiaoqi Jiao, ♦Mengjiao Peng and Yong Zhou.* East China Normal University

11:15 Estimation and Variable Selection for Single-Index Models with Right Censored Data via Distance Covariance

*Xiaohui Yuan.*

11:40 Floor Discussion.

#### **Session 23CHI118: Experimental Designs and their Applications**

Room: CHONGMING HALL

Organizer: Yongdao Zhou, Nankai University, China.

Chair: Jianhui Ning, Central China Normal University, China.

10:00 Revised Schematic Array

*Zuoluo Hao and ♦Yu Tang.* Soochow University

10:25 Group-Orthogonal Subsampling for Big Data Linear Mixed Models

*Jiaqing Zhu<sup>1</sup>, Lin Wang<sup>2</sup> and ♦Fasheng Sun<sup>1</sup>.* <sup>1</sup>NENU <sup>2</sup>PurdueU

10:50 Sequentially Weighted Uniform Designs

*Yao Xiao<sup>1</sup>, Shiqi Wang<sup>2</sup>, Hong Qin<sup>1</sup> and ♦Jianhui Ning<sup>2</sup>.* <sup>1</sup>Zhongnan University of Economics and Law <sup>2</sup>Central China Normal University

11:15 Uniform Design with Prior Information of Factors under Weighted Wrap-Around L2-Discrepancy

*Zujun Ou.* Jishou University

11:40 Floor Discussion.

#### **Session 23CHI136: New Statistical Learning for High-Dimensional Data**

Room: T-1

Organizer: Annie Qu, UC Irvine.

Chair: Yuxin Chen, University of Pennsylvania.

10:00 Transfer Learning for High-Dimensional Quantile Regression via Convolution Smoothing

<sup>♦</sup>*Yijiao Zhang and Zhongyi Zhu.* Fudan University

10:25 Center-Augmented  $\ell_2$ -Type Regularization for Subgroup Learning

*Huazhen Lin.* Southwestern University of Finance and Economics

10:50 Transfer Learning for High-Dimensional Quantile Regression via Convolution Smoothing

<sup>♦</sup>*Yijiao Zhang and Zhongyi Zhu.* Fudan University

11:15 Learning Individualized Minimal Clinically Important Difference (Imcid) from High-Dimensional Data  
*Jiwei Zhao.* University of Wisconsin Madison

11:40 Deflated Heteropca: Overcoming the Curse of Ill-Conditioning in Heteroskedastic Pca  
*Yuchen Zhou and \*Yuxin Chen.* University of Pennsylvania  
 Floor Discussion.

### Session 23CHI165: Recent Advances in Scalable Bayesian Inference

Room: LUWAN HALL

Organizer: Xin Tong, National University of Singapore, Singapore.  
 Chair: Cheng Li, National University of Singapore, Singapore.

10:00 Learnable Topological Features for Phylogenetic Inference  
*Cheng Zhang.* Peking University

10:25 Bayesian Fixed-Domain Asymptotics for Covariance Parameters in Spatial Gaussian Process Regression Models  
 ♦*Cheng Li<sup>1</sup>, Saifei Sun<sup>1</sup> and Yichen Zhu<sup>2</sup>.* <sup>1</sup>National University of Singapore <sup>2</sup>Duke University

10:50 Catalytic Priors: using Synthetic Data to Specify Prior Distributions in Bayesian Analysis  
 ♦*Dongming Huang<sup>1</sup>, Feicheng Wang<sup>2</sup>, Donald Rubin<sup>2</sup> and Samuel Kou<sup>2</sup>.* <sup>1</sup>National University of Singapore <sup>2</sup>Harvard University

11:15 Sampling with Constraints using Variational Methods  
*Xin Tong.* National University of Singapore

11:40 Floor Discussion.

### Session 23CHI17: Statistical Methods for Complex Longitudinal and Survival Data

Room: JINGAN HALL

Organizer: Gang Li, UCLA.  
 Chair: Gang Li, UCLA.

10:00 Efficient Algorithms for Survival Data with Multiple Outcomes using the Frailty Model  
*Xifan Huang, Jinfeng Xu and \*Yunpeng Zhou.*

10:25 Semiparametric Regression Analysis of Interval-Censored Multi-State Data with an Absorbing State  
 ♦*Yu Gu, Donglin Zeng and Danyu Lin.* University of North Carolina at Chapel Hill

10:50 Sure Joint Screening for High Dimensional Cox's Proportional Hazards Model under the Case-Cohort Design  
 ♦*Yi Liu<sup>1</sup> and Gang Li<sup>2</sup>.* <sup>1</sup>Ocean University of China  
<sup>2</sup>University of California at Los Angeles

11:15 Federated Survival Analysis via Data Augmentation using Multi-Task Variational Autoencoder  
*Hong Wang.* Central South University

11:40 Floor Discussion.

### Session 23CHI2: Recent Developments in Network Analysis and High Dimensional Data

Room: YANGPU HALL

Organizer: Wanjie Wang, National University of Singapore.  
 Chair: Jialiang Li, National University of Singapore.

10:00 Higher-Order Accurate Two-Sample Network Inference and Network Hashing

*Meijia Shao<sup>1</sup>, Dong Xia<sup>2</sup>, \*Yuan Zhang<sup>1</sup>, Qiong Wu<sup>3</sup> and Shuo Chen<sup>4</sup>.* <sup>1</sup>The Ohio State University <sup>2</sup>Hong Kong University of Science and Technology <sup>3</sup>University of Pennsylvania <sup>4</sup>University of Maryland, Baltimore

10:25 Ranking Inferences Based on the Top Choice of Multiway Comparisons

*Jianqing Fan<sup>1</sup>, Zhipeng Lou<sup>1</sup>, \*Weichen Wang<sup>2</sup> and Mengxin Yu<sup>1</sup>.* <sup>1</sup>Princeton University <sup>2</sup>The University of Hong Kong

10:50 Spectral Analysis on Networks with Attributes  
*Wanjie Wang.* National University of Singapore

11:15 Snr Estimation under High-Dimensional Linear Models  
*Xiaohan Hu and \*Xiaodong Li.* UC Davis

11:40 Floor Discussion.

### Session 23CHI4: Hot and Special Biostatistical Considerations in HA Guidelines

Room: CHANGNING HALL

Organizer: Qian Jiang, Department of Biostatistics and Programming, China, Sanofi, Inc..  
 Chair: Qian Jiang, Department of Biostatistics and Programming, China, Sanofi, Inc..

10:00 Missing Data Handling for Time-to-Event Data under the Framework of Estimand  
*Bin Yao and \*Lei Wang.* Beigene

10:25 Covariate-Adjusted Analysis Targeting Marginal Estimand for Binary and Timetoevent Outcomes: Considerations and Examples  
*Xin Zhang.* Pfizer Inc.

10:50 Concentration-Qtc Analysis to Support Ich E14 with a Real Case  
*Kong Xin.*

11:15 Immuno-Bridging in Vaccine Development

♦*Iris Sun<sup>1</sup> and Lanfang Xia<sup>2</sup>.* <sup>1</sup>Department of Biostatistics and programming, China, Sanofi, Inc. <sup>2</sup>PPD, part of Thermo Fisher Scientific

11:40 Floor Discussion.

### Session 23CHI59: Recent Advances in Survey Statistics and Data Science

Room: NANSHI HALL

Organizer: Zhengyuan Zhu, Iowa State University.  
 Chair: Cindy Yu, Iowa State University.

10:00 Loss Distribution of Arbitrary Credit Portfolios  
*Yimin Yang.* Loyal Trust Bank

10:25 Variable Selection on Survey Data with Missing Values  
*Yang Li<sup>1</sup>, Haoyu Yang<sup>1</sup>, Haochen Yu<sup>1</sup>, Hanwen Huang<sup>2</sup> and Ye Shen<sup>2</sup>.* <sup>1</sup>Renmin University of China <sup>2</sup>University of Georgia

- 10:50 Robust Personalized Federated Learning with Sparse Penalization  
*Weidong Liu<sup>1</sup>, Xiaojun Mao<sup>1</sup>, ♦Xiaofei Zhang<sup>2</sup> and Xin Zhang<sup>3</sup>. <sup>1</sup>Shanghai Jiao Tong University <sup>2</sup>Zhongnan University of Economics and Law <sup>3</sup>Iowa State University*
- 11:15 Correlated Quantization for Distributed Mean Estimation — a Sampler's Perspective  
*Xiaojun Mao<sup>1</sup>, ♦Hengfang Wang<sup>2</sup> and Xiaofei Zhang<sup>3</sup>. <sup>1</sup>Shanghai Jiao Tong University <sup>2</sup>Fujian Normal University <sup>3</sup>Zhongnan University of Economics and Law*
- 11:40 Floor Discussion.

### Session 23CHI62: Recent Developments in Change Point Analysis and Related Topics

Room: JIADING HALL

Organizer: Wei Ning, Bowling Green State University.  
 Chair: Weizhong Tian, Department of Mathematics, Shenzhen Technology University.

- 10:00 Generalized Fiducial Inference for Gev Change-Point Model  
 ♦*Xia Cai, Yaru Qiao and Shanshan Li.* Hebei University of Science and Technology
- 10:25 Integrative Learning of Linear Non-Gaussian Directed Acyclic Graphs  
 ♦*Xuanyu Li<sup>1</sup> and Qingzhao Zhang<sup>2</sup>.* <sup>1</sup>University of Chinese Academy of Sciences <sup>2</sup>Xiamen University
- 10:50 Confidence Intervals for Heterogeneity in Meta-Analysis of the Rare Binary Events Based on Empirical Likelihood-Type Methods  
*Sha Li<sup>1</sup>, ♦Weizhong Tian<sup>2</sup>, Ximin Li<sup>1</sup> and Wei Ning<sup>3</sup>. <sup>1</sup>School of Mathematics and Statistics, Qingdao University <sup>2</sup>College of Big Data and Internet, Shenzhen Technology University, Shenzhen <sup>3</sup>Department of Mathematics and Statistics, Bowling Green State University*
- 11:15 Monitoring Sequential Structural Changes in Penalized High-Dimensional Linear Models  
*Wei Ning.* Bowling Green State University
- 11:40 Floor Discussion.

### Session 23CHI68: Mendelian Randomization: Causal Inference and Beyond

Room: NANHUI HALL

Organizer: Xiaofeng Zhu, Department of Population and Quantitative Health Sciences, Case Western Reserve University.  
 Chair: Xiaofeng Zhu, Department of Population and Quantitative Health Sciences, Case Western Reserve University.

- 10:00 Adjusting for Genetic Confounders in Transcriptome-Wide Association Studies Leads to Reliable Detection of Causal Genes  
*Siming Zhao<sup>1</sup>, Wesley Crouse<sup>2</sup>, ♦Sheng Qian<sup>2</sup>, Kaixuan Luo<sup>2</sup>, Matthew Stephens<sup>3</sup> and Xin He<sup>2</sup>. <sup>1</sup>Department of Biomedical Data Science, Dartmouth Cancer Center, Dartmouth College <sup>2</sup>Department of Human Genetics, University of Chicago <sup>3</sup>Department of Human Genetics, Department of Statistics, University of Chicago*

- 10:25 Likelihood-Based Mendelian Randomization Analysis with Automated Instrument Selection and Horizontal Pleiotropic Modeling  
*Xiang Zhou.* University of Michigan
- 10:50 Identifying Gene by Environment Interactions, a Framework of Mr Approach  
*Xiaofeng Zhu.* Case Western Reserve University
- 11:15 Robust Multivariable Mendelian Randomization Based on Constrained Maximum Likelihood  
*Zhaotong Lin, Haoran Xue and ♦Wei Pan.* University of Minnesota
- 11:40 Floor Discussion.

### Session 23CHI69: Recent Advances in Adaptive Clinical Trial Design

Room: SONGJIANG HALL

Organizer: Samuel Wu, University of Florida.  
 Chair: Guogen Shan, University of Florida.

- 10:00 Adaptive Promising Zone Two-Stage Design for a Trial with Binary Endpoint  
*Guogen Shan.* University of Florida
- 10:25 Floor Discussion.

### Session 23CHI76: Statistical Methods and Application

Room: JINSHAN HALL

Organizer: Haitao Zheng, Southwest Jiaotong University.  
 Chair: Qing Cheng, Southwest University of Finance and Economics.

- 10:00 Additive Autoregressive Models for Matrix Valued Time Series  
*Hong-Fan Zhang.* Southwest Jiaotong University
- 10:25 A Stable and Adaptive Polygenic Signal Detection Method Based on Repeated Sample Splitting  
 ♦*Yanyan Zhao<sup>1</sup> and Lei Sun<sup>2</sup>.* <sup>1</sup>Shandong University <sup>2</sup>University of Toronto
- 10:50 A Bayesian Latent Subgroup Phase i/I Platform Design to Co-Develop Optimal Biological Doses for Multiple Indications  
 ♦*Rongji Mu<sup>1</sup>, Xiaojiang Zhan<sup>2</sup> and Ying Yuan<sup>3</sup>.* <sup>1</sup>Shanghai Jiao Tong University <sup>2</sup>Servier Pharmaceuticals <sup>3</sup>The University of Texas MD Anderson Cancer Center
- 11:15 Jackknife Empirical Likelihood for the Lower Mean Ratio  
 ♦*Lei Huang, Li Zhang and Yichuan Zhao.*
- 11:40 Discussant: Hongfan Zhang, Southwest Jiaotong University  
 Floor Discussion.

### Session 23CHI81: Statistical Inferences for Complex Data

Room: PUTUO HALL

Organizer: Li Cai, Zhejiang Gongshang University.  
 Chair: Li Cai, Zhejiang Gongshang University.

- 10:00 Inference for Arma Time Series with Mildly-Varying Trend  
 ♦*Yinghuai Yi, Zening Song and Lijian Yang.* Tsinghua University

- 10:25 Robust Regression using Probabilistically Linked Data  
*Raymond Chambers<sup>1</sup>, Enrico Fabrizi<sup>2</sup>, Maria Ranalli<sup>3</sup>, Nichola Salvati<sup>4</sup> and \*Suojin Wang<sup>5</sup>.* <sup>1</sup>University of Wollongong <sup>2</sup>Università Cattolica del Sacro Cuore <sup>3</sup>Università degli Studi di Perugia <sup>4</sup>Università di Pisa <sup>5</sup>Texas A&M University
- 10:50 Jackknife Empirical Likelihood for the Mean Difference of Two Zero-Inflated Skewed Populations  
*Faysal Satter<sup>1</sup> and \*Yichuan Zhao<sup>2</sup>.* <sup>1</sup>Lowe's Companies <sup>2</sup>Georgia State University
- 11:15 Testing Linearity in Semi-Functional Partially Linear Regression Models  
*\*Yongzhen Feng<sup>1</sup>, Jie Li<sup>2</sup> and Xiaojun Song<sup>3</sup>.* <sup>1</sup>Tsinghua University <sup>2</sup>Renmin University of China <sup>3</sup>Peking University
- 11:40 Floor Discussion.

### Session 23CHI88: Intelligent Algorithms and Data Mining

Room: FENGXIAN HALL

Organizer: Chang Li, Shandong University of Finance and Economics.

Chair: Chang Li, Shandong University of Finance and Economics.

10:00 Floor Discussion.

**July 1, 13:00-14:40**

### Session 23CHI129: Optimal Designs

Room: LONGFENG HALL-III

Organizer: Min-Qian Liu, Nankai University.

Chair: Yubin Tian, Beijing Institute of Technology.

13:00 Optimal Row-Column Designs

\*Zheng Zhou and Yongdao Zhou. Nankai University

13:25 Design Admissibility and De La Garza Phenomenon in Multi-Factor Experiments

*Holger Dette, \*Xin Liu<sup>1</sup> and Rong-Xian Yue<sup>2</sup>.* <sup>1</sup>Donghua University <sup>2</sup>Shanghai Normal University

13:50 Noncircular Designs for Controlling Border Effects under the Interference Model: Gain with No Pain

\*Xiangshun Kong<sup>1</sup>, Jie Fu<sup>1</sup> and Wei Zheng<sup>2</sup>. <sup>1</sup>Beijing institute of technology <sup>2</sup>University of Tennessee. Knoxville

14:15 Equivalence Theorems for c and Da-Optimality for Linear Mixed Effects Models with Applications to Multi-Treatment Group Assignments in Healthcare

*Xin Liu<sup>1</sup>, \*Rong-Xian Yue<sup>2</sup> and Weng Kee Wong<sup>3</sup>.* <sup>1</sup>College of Science, Donghua University, Shanghai 201620, China

<sup>2</sup>Department of Mathematics, Shanghai Normal University, Shanghai 200234, China <sup>3</sup>Department of Biostatistics, University of California, Los Angeles, CA 90095-1772, USA

14:40 Floor Discussion.

### Session 23CHI131: Advances in Bioinformatics, Data Science, and Clinical Trial Design

Room: T-1

Organizer: Hongjian Zhu, AbbVie Inc..

Chair: Hongjian Zhu, AbbVie Inc..

- 13:00 A Novel Transcriptional Risk Score for Risk Prediction of Complex Human Diseases  
*\*Nayang Shan<sup>1</sup>, Yuhuan Xie<sup>2</sup>, Shuang Song<sup>3</sup>, Wei Jiang<sup>2</sup>, Zuoheng Wang<sup>2</sup> and Lin Hou<sup>3</sup>.* <sup>1</sup>Capital University of Economics and Business <sup>2</sup>Yale School of Public Health <sup>3</sup>Tsinghua University

13:25 Mendelian Randomization for Causal Inference Accounting for Pleiotropy and Sample Structure using Genome-Wide Summary Statistics

\*Xianghong Hu<sup>1</sup>, Jia Zhao<sup>1</sup>, Zhixiang Lin<sup>2</sup>, Yang Wang<sup>1</sup>, Heng Peng<sup>3</sup>, Hongyu Zhao<sup>4</sup>, Xiang Wan<sup>5</sup> and Can Yang<sup>1</sup>.

<sup>1</sup>The Hong Kong University of Science and Technology

<sup>2</sup>The Chinese University of Hong Kong <sup>3</sup>Hong Kong Baptist University <sup>4</sup>Yale School of Public Health <sup>5</sup>Shen Zhen Research Institute of Big Data

- 13:50 Seamless Phase II/III Clinical Trials with Covariate Adaptive Randomization  
*Wei Ma<sup>1</sup>, Mengxi Wang<sup>2</sup> and \*Hongjian Zhu<sup>3</sup>.* <sup>1</sup>China Renmin University <sup>2</sup>UTHealth <sup>3</sup>AbbVie Inc.

14:15 Floor Discussion.

### Session 23CHI134: Message Passing and Differential Privacy

Room: XUHUI HALL

Organizer: Annie Qu, UC Irvine.

Chair: Yuting Wei, Department of Statistics and Data Science, University of Pennsylvania.

13:00 On Reference Panel-Based Regularized Estimators in High-Dimensional Sparsity-Free Genetic Data Prediction

\*Buxin Su<sup>1</sup>, Qiang Sun<sup>2</sup>, Xiao Chen Yang<sup>3</sup> and Bingxin Zhao<sup>1</sup>. <sup>1</sup>University of Pennsylvania <sup>2</sup>University of Toronto <sup>3</sup>Purdue University

13:25 Identification and Estimation of Causal Inference with Confounders Missing not at Random

\*Jian Sun and Bo Fu. 复旦大学

13:50 Differential Private Data Release for Mixed-Type Data via Latent Factor Models

\*Yanqing Zhang<sup>1</sup>, Qi Xu<sup>2</sup>, Niansheng Tang<sup>1</sup> and Annie Qu<sup>2</sup>.

<sup>1</sup>Yunnan University <sup>2</sup>University of California Irvine

14:15 The Distribution of Lasso and Its Applications: Arbitrary Covariance

*Yuting Wei.* University of Pennsylvania

14:40 Floor Discussion.

### Session 23CHI144: Applications of Bayesian Methods in Educational Statistics

Room: YANGPU HALL

Organizer: Ming-Hui Chen, University of Connecticut.

Chair: Zhihua Ma, Shenzhen University.

13:00 Bayesian Model Assessment in Joint Modeling of Response and Response Time Data under the Generalized Semi-Parametric Model

\*Fang Liu<sup>1</sup> and Ming-Hui Chen<sup>2</sup>. <sup>1</sup>Soochow University <sup>2</sup>University of Connecticut

- 13:25 Bayesian Inference for Multidimensional Irt Models with Flexible Distributions  
♦*Xue Zhang*<sup>1</sup>, *Chun Wang*<sup>2</sup> and *David Weiss*<sup>3</sup>. <sup>1</sup>Northeast Normal University <sup>2</sup>University of Washington <sup>3</sup>University of Minnesota
- 13:50 A Sequential Bayesian Changepoint Detection Procedure for Aberrant Behaviors in Computerized Testing  
♦*Jing Lu*<sup>1</sup>, *Chun Wang*<sup>2</sup>, *Jiwei Zhang*<sup>1</sup> and *Xue Wang*<sup>1</sup>.  
<sup>1</sup>Northeast Normal University <sup>2</sup>University of Washington
- 14:15 A Sequential Bayesian Changepoint Detection Procedure for Aberrant Behaviors in Computerized Testing  
♦*Jing Lu*<sup>1</sup>, *Chun Wang*<sup>2</sup>, *Jiwei Zhang*<sup>1</sup> and *Xue Wang*<sup>1</sup>.  
<sup>1</sup>Northeast Normal University <sup>2</sup>University of Washington
- 14:40 Variational Bayesian Algorithm in Dina Model  
*Juntao Want.*  
Floor Discussion.

### Session 23CHI155: Statistical Frontiers in Spatial, Single-Cell, and Single-Molecule Multi-Omics Data

Room: PUTUO HALL

Organizer: Shu Yang, North Carolina State University.

Chair: Wang Miao, Peking University.

- 13:00 Distributed Semi-Supervised Sparse Statistical Inference  
*Jiyuan Tu*<sup>1</sup>, *Weidong Liu*<sup>2</sup>, ♦*Xiaojun Mao*<sup>2</sup> and *Mingyue Xu*<sup>3</sup>. <sup>1</sup>Shanghai University of Finance and Economics <sup>2</sup>Shanghai Jiao Tong University <sup>3</sup>Columbia University
- 13:25 Functional Calibration under Non-Probability Survey Sampling  
♦*Zhonglei Wang*<sup>1</sup>, *Xiaojun Mao*<sup>2</sup> and *Jae Kwang Kim*<sup>3</sup>.  
<sup>1</sup>Xiamen University <sup>2</sup>Shanghai Jiao Tong University <sup>3</sup>Iowa State University
- 13:50 Semiparametric Regression Based on Quadratic Inference Function for Multivariate Failure Time Data with Auxiliary Information  
*Feifei Yan*, *Lin Zhu*, ♦*Yanyan Liu*, *Jianwen Cai* and *Haibo Zhou*.
- 14:15 Floor Discussion.

### Session 23CHI166: Recent Developments in Recurrent Event Data and Panel Count Data Analysis

Room: LUWAN HALL

Organizer: Ni Li, School of Mathematics and Statistics, Hainan Normal University, Haikou, China.

Chair: Weiwei Wang, School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, China.

- 13:00 Estimation and Inference for Fixed Center Effects on Panel Count Data  
♦*Weiwei Wang*<sup>1</sup>, *Yijun Wang*<sup>1</sup> and *Xiaobing Zhao*<sup>2</sup>.  
<sup>1</sup>Zhejiang Gongshang University <sup>2</sup>Zhejiang University of Finance and Economics
- 13:25 Regression Analysis of Mixed Panel Count Data with Dependent Observation Processes  
♦*Lei Ge*<sup>1</sup>, *Jahee Choi*<sup>2</sup>, *Hui Zhao*<sup>3</sup>, *Yang Li*<sup>4</sup> and *Jian-guo Sun*<sup>5</sup>. <sup>1</sup>Department of Biostatistics and Health Data

Science, Indiana University School of Medicine, Indianapolis, IN, USA; <sup>2</sup>Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, USA <sup>3</sup>School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, People's Republic of China <sup>4</sup>Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, USA <sup>5</sup>Department of Statistics, University of Missouri, Columbia, MO, USA

- 13:50 An Ipw Additive Model for Panel Count Data with Dependent Observation Process and Terminal Event  
*Ni Li*. Hainan Normal University
- 14:15 A Double Exponential Gamma-Frailty Model for Clustered Survival Data  
♦*Mengqi Xie*<sup>1</sup>, *Jie Zhou*<sup>1</sup> and *Lei Liu*<sup>2</sup>. <sup>1</sup>School of Mathematics, Capital Normal University, Beijing 100048, China <sup>2</sup>Division of Biostatistics, Washington University in St. Louis, MO 63110, USA
- 14:40 Floor Discussion.

### Session 23CHI26: Current Topics in Biostatistics I

Room: JIADING HALL

Organizer: Somnath Datta, University of Florida.

Chair: Peihua Qiu, University of Florida.

- 13:00 Bootstrap Averaging Reduces the Mean Square Error of Kinetic Maps Recovered from Pet Data  
♦*Fengyun Gu*<sup>1</sup>, *Finbarr O'sullivan*<sup>2</sup> and *Qi Wu*<sup>2</sup>. <sup>1</sup>North China Electric Power University <sup>2</sup>University College Cork
- 13:25 A Bayesian Non-Parametric Approach for Causal Mediation with a Post-Treatment Confounder  
*Woojung Bae* and ♦*Michael Daniels*. University of Florida
- 13:50 Transparent Sequential Learning for Monitoring Sequential Processes  
♦*Peihua Qiu* and *Xiulin Xie*. University of Florida
- 14:15 Floor Discussion.

### Session 23CHI34: Optimization in Statistics

Room: SONGJIANG HALL

Organizer: Xueqin Wang, University of Science and Technology of China.

Chair: Xueqin Wang, University of Science and Technology of China.

- 13:00 Fast Consistent Best-Subset Selection in Generalized Linear Models  
♦*Junxian Zhu*<sup>1</sup>, *Jin Zhu*<sup>2</sup>, *Borui Tang*<sup>3</sup>, *Xuanyu Chen*<sup>2</sup> and *Xueqin Wang*<sup>3</sup>. <sup>1</sup>National University of Singapore <sup>2</sup>Sun Yat-Sen University <sup>3</sup>University of Science and Technology of China
- 13:25 High Dimensional Portfolio Selection with Cardinality Constraints  
♦*Yifeng Guo*<sup>1</sup>, *Jin-Hong Du*<sup>2</sup> and *Xueqin Wang*<sup>3</sup>. <sup>1</sup>The University of Hongkong <sup>2</sup>Carnegie Mellon University <sup>3</sup>University of Science and Technology of China

- 13:50 Directed Community Detection with Network Embedding  
♦*Jingnan Zhang<sup>1</sup>, Xin He<sup>2</sup> and Junhui Wang<sup>3</sup>*. <sup>1</sup>University of Science and Technology of China <sup>2</sup>Shanghai University of Finance and Economics, China <sup>3</sup>The Chinese University of Hong Kong
- 14:15 Abess: a Fast Best-Subset Selection Library in Python and r  
*Jin Zhu*. Department of Statistical Science, Sun Yat-Sen University
- 14:40 Discussant: Yifeng Guo, The University of Hongkong  
Floor Discussion.

### Session 23CHI44: New Developments in Large-Scale and High-Dimensional Inference

Room: CHONGMING HALL  
Organizer: Wenguang Sun, Zhejiang University.  
Chair: Wenguang Sun, Zhejiang University.

- 13:00 Statistically Guided Divide-and-Conquer for Sparse Factorization of Large Matrix  
*Kun Chen, Ruipeng Dong, Wanwan Xu and ♦Zemin Zheng*.
- 13:25 Asymptotics of the Spatial-Sign Based Estimators of Location and Scatter in High-Dimension  
*Qinwen Wang*. FUDAN UNIVERSITY
- 13:50 Ranking and Selection in Large-Scale Inference of Heteroscedastic Units  
♦*Bowen Gang<sup>1</sup>, Luella Fu<sup>2</sup>, Wenguang Sun<sup>3</sup> and Gareth James James<sup>4</sup>*. <sup>1</sup>Fudan University <sup>2</sup>San Francisco State University <sup>3</sup>Zhejiang University <sup>4</sup>Emory University
- 14:15 Multi-Dimensional Domain Generalization with Low-Rank Structures  
♦*Sai Li<sup>1</sup> and Linjun Zhang<sup>2</sup>*. <sup>1</sup>Renmin University of China <sup>2</sup>Rutgers University
- 14:40 Floor Discussion.

### Session 23CHI49: New Development of Statistical Methods for Genomic, Epigenomic, and Microbiome Data

Room: CHANGNING HALL  
Organizer: Yijuan Hu, Emory University.  
Chair: Yijuan Hu, Emory University.

- 13:00 Pan-Cancer Analysis of Pathway-Based Gene Expression Pattern at the Individual Level Reveals Biomarkers of Clinical Prognosis  
*Zhaohui Qin*. Emory University
- 13:25 Microbial Risk Score for Capturing Microbial Characteristics, Integrating Multi-Omics Data, and Predicting Disease Risk  
♦*Chan Wang, Leopoldo Segal, Jiyuan Hu, Boyan Zhou, Richard Hayes, Jiyoung Ahn and Huilin Li*. New York University Grossman School of Medicine
- 13:50 Microbiome Composition-on-Composition Regression Analysis  
*Xiang Zhan*. Peking University
- 14:15 Floor Discussion.

### Session 23CHI6: Case Studies in Clinical Trial Design, Analysis and Result Interpretation

Room: LONGFENG HALL-I  
Organizer: Qian Jiang, Department of Biostatistics and Programming, China, Sanofi, Inc..  
Chair: Qian Jiang, Department of Biostatistics and Programming, China, Sanofi, Inc..

- 13:00 Leverage Registry Database and Natural History Study to Facilitate Full Approval—a Real Case Study  
*Grace Lin*. Sanofi
- 13:25 Propensity Score Matched External Control: a Use Case in Rare Disease Pediatric Clinical Trial Study  
*Ning Li*. Sanofi
- 13:50 Assessing the Probability of Clinical Trial Success via Modeling and Simulation: a Case Study  
*Michael Lee*. Harbour BioMed, Inc.
- 14:15 Floor Discussion.

### Session 23CHI71: Addressing Challenges in Time-to-Event Data

Room: JINSHAN HALL  
Organizer: Qihuang Zhang, McGill University.  
Chair: Qihuang Zhang, McGill University.

- 13:00 Learning Optimal Individualized Treatment Rules with Interval-Censored Data  
*Yichi Zhang<sup>1</sup> and ♦Yinghao Pan<sup>2</sup>*. <sup>1</sup>Yale University <sup>2</sup>University of North Carolina at Charlotte
- 13:25 A Semiparametric Approach to Develop Well-Calibrated Risk Assessment Models in Calculating Lifetime Risk for Breast Cancer  
♦*Yaqi Cao<sup>1</sup>, Ying Yang<sup>2</sup> and Jinbo Chen<sup>3</sup>*. <sup>1</sup>Minzu University of China <sup>2</sup>Tsinghua University <sup>3</sup>University of Pennsylvania
- 13:50 Cox Model with Left-Truncation and Auxiliary Outcomes  
♦*Yidan Shi and Sharon X. Xie*. University of Pennsylvania Perelman School of Medicine
- 14:15 Boosting Method for Length-Biased and Interval-Censored Survival Data Subject to High-Dimensional Error-Prone Covariates  
♦*Li-Pang Chen and Bangxu Qiu*. National Chengchi University
- 14:40 Floor Discussion.

### Session 23CHI72: How to Dissect and Understand Diverse High-Throughput Data by Novel Statistical and Computational Methods?

Room: BAOSHAN HALL  
Organizer: Fangda Song, The Chinese University of Hong Kong, Shenzhen.  
Chair: Fangda Song, The Chinese University of Hong Kong, Shenzhen.

- 13:00 Deconvolution of Bulk Rna-Seq Reveals Cell-Type Specificity Mechanism in Alzheimer's Disease  
*Yun Li*. University of North Carolina at Chapel Hill

- 13:25 A Prism Vote Method for Risk Prediction of Traits in Genotype Data of Multi-Population  
♦Xiaoxuan Xia, Ming Hin Ng, Lin Hou and Yingying Wei.
- 13:50 Identification of Cell-Type-Specific Spatially Variable Genes Accounting for Excess Zeros  
Xiangyu Luo. Renmin University of China
- 14:15 Gene-Environment Interaction Analysis via Deep Learning  
Shuni Wu<sup>1</sup>, Yaqing Xu<sup>2</sup>, Qingzhao Zhang<sup>3</sup> and ♦Shuangge Ma<sup>4</sup>. <sup>1</sup>The Wang Yanan Institute for Studies in Economics, Xiamen University <sup>2</sup>Shanghai Jiao Tong University School of Medicine <sup>3</sup>Department of Statistics and Data Science, School of Economics and Fujian Key Lab of Statistics, Xiamen University <sup>4</sup>Department of Biostatistics, Yale School of Public Health
- 14:40 Floor Discussion.

### Session 23CHI74: Recent Advances in Functional Data Analysis

Room: JINGAN HALL

Organizer: Wenlin Dai, Renmin University of China.  
Chair: Wenlin Dai, Renmin University of China.

- 13:00 A Unified Analysis of Multi-Task Functional Linear Regression Models with Manifold Constraint and Composite Quadratic Penalty  
Shiyuan He. Renmin University
- 13:25 Exponential-Family Principal Component Analysis of Two-Dimensional Functional Data with Serial Correlation  
Shirun Shen<sup>1</sup>, ♦Kejun He<sup>1</sup>, Bohai Zhang<sup>2</sup> and Lan Zhou<sup>3</sup>. <sup>1</sup>Renmin University of China <sup>2</sup>Nankai University <sup>3</sup>Texas A&M University
- 13:50 Latent Group Detection in Functional Partially Linear Regression Models  
♦Wu Wang<sup>1</sup>, Ying Sun<sup>2</sup> and Huixia Wang<sup>3</sup>. <sup>1</sup>Center for Applied Statistics and School of Statistics, Renmin University of China <sup>2</sup>Statistics Program, King Abdullah University of Science and Technology <sup>3</sup>Department of Statistics, The George Washington University, Washington
- 14:15 Global Depths for Irregularly Observed Multivariate Functional Data  
Zhuo Qu<sup>1</sup>, ♦Wenlin Dai<sup>2</sup> and Marc G. Genton<sup>1</sup>. <sup>1</sup>KAUST <sup>2</sup>Renmin University of China
- 14:40 Floor Discussion.

### Session 23CHI75: Recent Statistical Advances for Complex Multi-Omics Data Analysis

Room: NANSHI HALL

Organizer: Yuehua Cui, Michigan State University.  
Chair: Yuehua Cui, Michigan State University.

- 13:00 Kernel-Based Multi-Omics Data Integration for Cancer Subtyping  
Hongyan Cao. Shanxi Medical University

- 13:25 Multi-Omics Data Integration with Multi-View Learning via Composed Tensors  
♦Xu Liu<sup>1</sup>, Yiming Lliu<sup>1</sup>, Xiao Zhang<sup>1</sup>, Jian Huang<sup>2</sup>, Yuehua Cui<sup>3</sup> and Xingjie Shi<sup>4</sup>. <sup>1</sup>Shanghai University of Finance and Economics <sup>2</sup>The Hong Kong Polytechnic University <sup>3</sup>Michigan State University <sup>4</sup>East China Normal University
- 13:50 Leveraging Trans-Ethnic Genetic Risk Scores to Improve Association Power for Complex Traits in Underrepresented Populations  
Haojie Lu, Shuo Zhang, Zhou Jiang and ♦Ping Zeng. Xuzhou Medical University
- 14:15 An Adaptive Set-Based Testing Framework for High-Dimensional Association Studies  
Haitao Yang.
- 14:40 Floor Discussion.

### Session 23CHI80: Some Advances in Analysis of High-Dimensional Complex Data

Room: FENGXIAN HALL

Organizer: Wensheng Zhu, Northeast Normal University.  
Chair: Wensheng Zhu, Northeast Normal University.

- 13:00 Distribution Estimation of Contaminated Data via Dnn-Based Mom-Gans  
Fang Xie<sup>1</sup>, Lihu Xu<sup>2</sup>, Qiuran Yao<sup>2</sup> and ♦Huiming Zhang<sup>3</sup>. <sup>1</sup>BNU-HKBU United International College <sup>2</sup>University of Macau <sup>3</sup>Beihang University
- 13:25 Large-Scale Spatial Multiple Testing via Shifted p-Values  
Pengyu Yan and ♦Pengfei Wang. Dongbei University of Finance and Economics
- 13:50 Adaptive Learning of Personalized Tuning Parameters for Feature Selection in Linear Models  
Bin Wang, ♦Xiaofei Wang and Jianhua Guo. Northeast Normal University
- 14:15 On Statistical Analysis of High-Dimensional Factor Models  
Junfan Mao, ♦Zhigen Gao, Bingyi Jing and Jianhua Guo.
- 14:40 Discussant: Zhigen Gao, Northeast Normal University  
Floor Discussion.

### Session 23CHI83: New Advances in Statistical Methods for Health-Related Data

Room: NANHUI HALL

Organizer: Tiejun Tong, Hong Kong Baptist University.  
Chair: Ke Yang, Beijing University of Technology.

- 13:00 Admissibility Condition for Combining the Dependent *p*-Values and Its Application for Meta-Analysis  
♦Ke Yang<sup>1</sup> and Tiejun Tong<sup>2</sup>. <sup>1</sup>Beijing University of Technology <sup>2</sup>Hong Kong Baptist University
- 13:25 Standardized Mean Difference without the Homoscedasticity Assumption  
♦Jiandong Shi<sup>1</sup>, Xiaochen Zhang<sup>2</sup> and Tiejun Tong<sup>3</sup>. <sup>1</sup>The Hong Kong University of Science and Technology <sup>2</sup>Shandong University <sup>3</sup>Hong Kong Baptist University

- 13:50 Sparse Convolved Rank Regression in High Dimensions  
 ♦Le Zhou<sup>1</sup>, Boxiang Wang<sup>2</sup> and Hui Zou<sup>3</sup>. <sup>1</sup>Hong Kong Baptist University <sup>2</sup>University of Iowa <sup>3</sup>University of Minnesota
- 14:15 A Nonparametric Mixed-Effects Mixture Model for Patterns of Clinical Measurements Associated with Covid-19  
 Xiaoran Ma<sup>1</sup>, Wensheng Guo<sup>2</sup>, Mengyang Gu<sup>1</sup>, Peter Kotanko<sup>3</sup>, Len Usvyat<sup>4</sup> and ♦Yuedong Wang<sup>1</sup>. <sup>1</sup>University of California - Santa Barbara <sup>2</sup>University of Pennsylvania <sup>3</sup>Renal Research Institute <sup>4</sup>Fresenius Medical Care
- 14:40 Floor Discussion.

### Session 23CHI95: Modern Statistical Process Control and Change-Point Problems I

Room: LONGFENG HALL-II

Organizer: Yajun Mei, H. Milton Stewart School of Industrial and Systems Engineering Georgia Institute of Technology; Dongdong Xiang, School of Statistics, East China Normal University.  
 Chair: Zhonghua Li, School of Statistics and Data Science, Nankai University.

- 13:00 Design of Ewma Control Chart with Time-Varying Smoothing Parameters Based on Bayesian Likelihood Ratio  
 ♦Baocai Guo and Pingye Gong. Zhejiang Gongshang University
- 13:25 A New Ewma Control Chart for Monitoring Variability  
 Dan Wang. Northwest University
- 13:50 Fault Diagnosis for High-Dimensional Data Streams under Two-Layers Mdr Controls  
 Dongdong Xiang. East China Normal Univesity
- 14:15 On-line Profile Monitoring for Surgical Outcomes using a Weighted Score Test  
 ♦Liu Liu, Xin Lai, Jian Zhang and Fugee Tsung.
- 14:40 Floor Discussion.

### July 1, 15:00-16:40

### Session 23CHI107: New Developments of Statistical Methods in Bio-Medical Studies

Room: CHANGNING HALL

Organizer: Ao Yuan, Georgetown University.  
 Chair: DianLiang Deng, University of Regina.

- 15:00 Estimation of Quantile Function for Gene Expression Trajectory under Multiple Biological Conditions  
 ♦Dianliang Deng and Qi Lyu. University of Regina
- 15:25 New Developments of Statistical Methods in Bio-Medical Studies  
 ♦Larry Tang and Ty Nguyen. University of Central Florida
- 15:50 Identification of Survival Relevant Genes with Measurement Error in Gene Expression Incorporated  
 Juan Xiong<sup>1</sup> and ♦Wenqing He<sup>2</sup>. <sup>1</sup>Shenzhen University <sup>2</sup>University of Western Ontario
- 16:15 Direct Estimation of Volume under the Roc Surface with Verification Bias  
 ♦Gengsheng Qin and Shuangfei Shi.

- 16:40 Floor Discussion.

### Session 23CHI111: Recent Development on High-Dimensional Data Analysis

Room: JINGAN HALL

Organizer: Jin-Ting Zhang, National University of Singapore, Singapore.  
 Chair: Wangli Xu, Remin University, Beijing, China.

- 15:00 An Adaptable Independence Test using Kernel Projection Criterion in High Dimensions  
 Yuexin Chen and ♦Wangli Xu.

- 15:25 Two-Sample Test for High-Dimensional Covariance Matrices: a Normal-Reference Approach  
 ♦Jin-Ting Zhang<sup>1</sup>, Jingyi Wang<sup>2</sup> and Tianming Zhu<sup>3</sup>.  
<sup>1</sup>Professor <sup>2</sup>PhD student <sup>3</sup>Assistant Professor
- 15:50 A Pairwise Hotelling Method for Testing High-Dimensional Mean Vectors  
 ♦Zongliang Hu<sup>1</sup>, Tiejun Tong<sup>2</sup> and Marc G. Genton<sup>3</sup>.  
<sup>1</sup>Shenzhen University <sup>2</sup>HongKong Baptist University <sup>3</sup>King Abdullah University of Science and Technology
- 16:15 Floor Discussion.

### Session 23CHI114: Topics on Statistical Study and Method

Room: NANHUI HALL

Organizer: Chunjie Wang, Changchun University of Technology.  
 Chair: Kai Yang, Changchun University of Technology.

- 15:00 Matrix Garch Model: Inference and Application  
 ♦Cheng Yu<sup>1</sup>, Dong Li<sup>1</sup>, Feiyu Jiang<sup>2</sup> and Ke Zhu<sup>3</sup>.  
<sup>1</sup>Tsinghua University <sup>2</sup>Fudan University <sup>3</sup>The University of Hong Kong
- 15:25 On Bivariate Threshold Poisson Integer-Valued Autoregressive Processes  
 ♦Kai Yang<sup>1</sup>, Yiwei Zhao<sup>1</sup>, Han Li<sup>2</sup> and Dehui Wang<sup>3</sup>.  
<sup>1</sup>Changchun University of Technology <sup>2</sup>Changchun University <sup>3</sup>Liaoning University
- 15:50 Semi-Supervised Learning for Two-Sample Comparison  
 Mengjiao Peng. East China Normal University
- 16:15 Floor Discussion.

### Session 23CHI15: Lifetime Data Analysis

Room: LONGFENG HALL-I

Organizer: Mei-Ling Ting Lee, University of Maryland.  
 Chair: Jiali Li, National University of Singapore.

- 15:00 Group Sequential Trial Design without Proportional Hazards Assumption  
 Yiming Chen<sup>1</sup>, John Lawrence<sup>1</sup> and ♦Mei-Ling Ting Lee<sup>2</sup>.  
<sup>1</sup>US Food & Drug Administration <sup>2</sup>University of Maryland
- 15:25 Semiparametric Estimation of Cause-Specific Regression Parameters and Cumulative Incidence Functions for Serial Gap Time Data with Recurrent Events and a Terminal Event  
 Shu-Hui Chang. National Taiwan University
- 15:50 Floor Discussion.

**Session 23CHI153: Advances in Regression Methods and Study Design**

Room: LONGFENG HALL-II

Organizer: Lei Liu, Washington University in St. Louis.

Chair: Li Wang, Abbvie.

15:00 Linearized Maximum Rank Correlation Estimation

*Guohao Shen<sup>1</sup>, Kani Chen<sup>2</sup>, Jian Huang<sup>1</sup> and ♦Yuanyuan Lin<sup>3</sup>.* <sup>1</sup>The Hong Kong Polytechnic University <sup>2</sup>Hong Kong University of Science and Technology <sup>3</sup>The Chinese University of Hong Kong

15:25 Complex Innovative Design Pilot Program and a Potential Proposal: Bayesian Predictive Platform Design for Proof of Concept and Dose Finding using Early and Late Endpoints  
*Li Wang.*

15:50 Functional Concurrent Hidden Markov Model

*Xiaoxiao Zhou and ♦Xinyuan Song.* The Chinese University of Hong Kong

16:15 Deep Learning for Regression Analysis of Interval-Censored Data

*Mingyue Du<sup>1</sup>, Qiang Wu<sup>2</sup>, Xingwei Tong<sup>2</sup> and ♦Xingqiu Zhao<sup>1</sup>.* <sup>1</sup>The Hong Kong Polytechnic University <sup>2</sup>Beijing Normal University

16:40 Floor Discussion.

**Session 23CHI169: Statistical Methods and Analysis for the Digital Asset Economy**

Room: T-1

Organizer: Jeffrey Chu, Renmin University of China, Beijing, China; American University of Sharjah, Dubai, United Arab Emirates.

Chair: Jeffrey Chu, Renmin University of China, Beijing, China; American University of Sharjah, Dubai, United Arab Emirates.

15:00 Price Divergence in Bitcoin Market

*Dehua Shen.* Nankai University

15:25 Do Clean and Dirty Cryptocurrencies Connect with Financial Assets Differently? the Role of Economic Policy Uncertainty

*Kun Duan<sup>1</sup>, ♦Yingying Huang<sup>2</sup>, Yanqi Zhao<sup>1</sup> and Andrew Urquhart<sup>3</sup>.* <sup>1</sup>Huazhong University of Science and Technology <sup>2</sup>Harbin Institute of Technology <sup>3</sup>University of Reading

15:50 An Analysis of the Return-volume Relationship in Decentralised Finance (Defi)

*♦Jeffrey Chu<sup>1</sup>, Stephen Chan<sup>2</sup> and Yuanyuan Zhang<sup>3</sup>.*

<sup>1</sup>Renmin University of China <sup>2</sup>American University of Sharjah <sup>3</sup>University of Manchester

16:15 The Effect of Central Bank Digital Currency Volatility on Supply Chain Management

*Shusheng Ding.*

16:40 Floor Discussion.

**Session 23CHI47: Advanced Statistical Methods in Omics Data Analysis**

Room: CHONGMING HALL

Organizer: Zhaohui Qin, Emory University.

Chair: Zhaohui Qin, Emory University.

15:00 Model-Based Spatial Reconstruction of Large-Scale Biomolecules via Bayesian Inference of a Hierarchical Spatial Model

*Chong Shen<sup>1</sup>, Shiyu Wang<sup>2</sup>, Zhaohui Qin<sup>2</sup> and ♦Ke Deng<sup>1</sup>.*  
<sup>1</sup>Tsinghua University <sup>2</sup>Emory University

15:25 Scemail: Universal and Source-Free Annotation Method for ScRNA-Seq Data with Novel Cell-Type Perception

*Hui Wan<sup>1</sup>, Liang Chen<sup>2</sup> and ♦Minghua Deng<sup>1</sup>.* <sup>1</sup>Peking University <sup>2</sup>Huawei Technologies Co.

15:50 Integrative Analysis of 16s Marker-Gene and Shotgun Metagenomic Sequencing Data Improves the Efficiency of Testing Hypotheses About the Microbiome

*Ye Yue, Glen Satten and ♦Yijuan Hu.* Emory University

16:15 Discussant: Zhaohui Qin, Emory University

16:40 Floor Discussion.

**Session 23CHI48: Emerging Development in Statistical Analyses for Multi-Omic Data**

Room: NANSHI HALL

Organizer: Yue Wang, University of Colorado Anschutz Medical Campus.

Chair: Yue Wang, University of Colorado Anschutz Medical Campus.

15:00 Bayesian Integrative Region Segmentation in Spatially Resolved Transcriptomic Studies

*♦Yinqiao Yan and Xiangyu Luo.* Renmin University of China

15:25 Enhancing the Study of Microbiome-Metabolome Interactions: a Transfer-Learning Approach for Precise Identification of Essential Microbes

*♦Yue Wang<sup>1</sup>, Lillian Li<sup>2</sup>, Chenglong Ye<sup>2</sup> and Tim Randolph<sup>3</sup>.*

<sup>1</sup>Colorado School of Public Health <sup>2</sup>University of Kentucky

<sup>3</sup>Fred Hutch Cancer Center

15:50 A Novel Transcriptome-Wide Association Study Method with Incorporating Multiple Annotations

*Han Wang and ♦Yan Dora Zhang.* The University of Hong Kong

16:15 Me-Bayes SI: Enhanced Bayesian Polygenic Risk Prediction Leveraging Information Across Multiple Ancestry Groups

*♦Jin Jin<sup>1</sup>, Jianan Zhan<sup>2</sup>, Jingning Zhang<sup>3</sup>, Ruzhang Zhao<sup>3</sup>, Jared O'connell<sup>2</sup>, Yunxuan Jiang<sup>2</sup>, Steven Buyske<sup>4</sup>, Genevieve Wojcik<sup>5</sup>, Haoyu Zhang<sup>6</sup> and Nilanjan Chatterjee<sup>7</sup>.* <sup>1</sup>Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health; Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania <sup>2</sup>23andMe, Inc. <sup>3</sup>Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

<sup>4</sup>Department of Statistics, Rutgers University <sup>5</sup>Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health <sup>6</sup>Division of Cancer Epidemiology and Genetics, National Cancer Institute <sup>7</sup>Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health; Department of Oncology, School of Medicine, Johns Hopkins University

16:40 Floor Discussion.

**Session 23CHI50: Recent Advances in Data Fusion and Integration with Real-World Applications in Healthcare**

Room: FENGXIAN HALL

Organizer: Kun Chen, University of Connecticut.

Chair: Liu Liu, Chengdu University of Technology.

15:00 Robust Machine Learner for Mean Potential Outcome with Information Integration from Auxiliary Data

♦*Chixiang Chen<sup>1</sup>, Shuo Chen, Zhenyao Ye, Xu Shi and Tianzhou Ma.* <sup>1</sup>university of maryland, school of medicine

15:25 Lossless One-Shot Distributed Linear Mixed Model for a Multi-Site International Study of Covid-19 Hospitalization Length of Stay

♦*Chongliang Luo<sup>1</sup>, Md. Nazmul Islam<sup>2</sup>, Jenna Reps<sup>3</sup>, Rui Duan<sup>4</sup>, Jiang Bian<sup>5</sup>, Talita Duarte-Salles<sup>6</sup>, Thomas Falconer<sup>7</sup>, Chungsoo Kim<sup>8</sup>, Hua Xu<sup>9</sup> and Yong Chen<sup>10</sup>.*<sup>1</sup>Washington University in St Louis <sup>2</sup>Optum Labs <sup>3</sup>JanssenResearch and Development LLC <sup>4</sup>Harvard T.H. Chan Schoolof Public Health, <sup>5</sup>University of Florida <sup>6</sup>Fundacio Institut

Universitari per a la recerca a l'Atencio Primaria de

Salut Jordi Gol i Gurina (IDIAPJGol) <sup>7</sup>Columbia University<sup>8</sup>Ajou University Graduate School of Medicine <sup>9</sup>The Universityof Texas Health Science Center at Houston <sup>10</sup>University

of Pennsylvania

15:50 Tree-Guided Rare Feature Selection and Logic Aggregation with Electronic Health Records Data

*Kun Chen.* University of Connecticut

16:15 Floor Discussion.

**Session 23CHI51: Big Data Analysis Based on Statistics and Machine Learning**

Room: JIADING HALL

Organizer: Xu Qin, School of Mathematical Sciences, University of Electronic Science and Technology of China.

Chair: Riquan Zhang, School of Statistics and Information, Shanghai University of International Business and Economics.

15:00 Statistical Analysis of Civil Aviation Big Data

*Riquan Zhang.* School of Statistics and Information, Shanghai University of International Business and Economics

15:25 Forecasting Stock Volatility and Value-at-Risk Based on Temporal Convolutional Networks

♦*Chunxia Zhang<sup>1</sup>, Jun Li<sup>1</sup>, Xingfang Huang<sup>2</sup>, Jiangshe Zhang<sup>1</sup> and Huachuan Huang<sup>1</sup>.* <sup>1</sup>Xi'an Jiaotong University <sup>2</sup>Nanjing Audit University

15:50 Deep Convolutional Neural Network with Feature Ensemble for Image Classification

*Yu Wang.* Shanxi University

16:15 An Improved SNE with Its Applications in Classification and Visualization

*Peilin Sun and ♦Xu Qin.* University of Electronic Science and Technology of China

16:40 Floor Discussion.

**Session 23CHI55: New Advancements in Analytical Methods using Multi-Source or Multi-Trial Data**

Room: LUWAN HALL

Organizer: Ming Wang, Case Western Reserve University.

Chair: Lijun Zhang, Case Western Reserve University.

15:00 Multiply-Robust Estimation of Causal Treatment Effect on a Binary Outcome with Integrated Information from Secondary Outcomes

*Chixiang Chen<sup>1</sup>, Shuo Chen<sup>1</sup>, Qi Long<sup>2</sup>, Sudeshna Das<sup>3</sup> and ♦Ming Wang<sup>4</sup>.* <sup>1</sup>University of Maryland <sup>2</sup>University of Pennsylvania <sup>3</sup>Harvard Medical School <sup>4</sup>Case Western Reserve University

15:25 Interpretable Image Segmentation Network Originated from Multi-Grid Variational Model

*Junying Meng<sup>1</sup>, ♦Weihong Guo<sup>2</sup>, Jun Liu<sup>1</sup> and Mingrui Yang<sup>3</sup>.* <sup>1</sup>Beijing Normal University <sup>2</sup>Case Western Reserve University <sup>3</sup>Cleveland Clinic

15:50 Statistical Concerns in Meta-Analysis of Safety Data

*Shouhao Zhou.* Penn State University

16:15 Semiparametric and Variable Marginal Effects with Debiased Machine Learning

♦*Chan Shen<sup>1</sup> and Roger Klein<sup>2</sup>.* <sup>1</sup>Penn State University <sup>2</sup>Rutgers University

16:40 Floor Discussion.

**Session 23CHI57: Emerging Problems and Solutions in Imaging Statistics**

Room: YANGPU HALL

Organizer: Jian Kang, University of Michigan.

Chair: Jian Kang, University of Michigan.

15:00 Survival Models with Medical Images as Predictors

*Zhangsheng Yu.* Shanghai Jiao Tong University

15:25 Latent Subgroup Identification in Image-on-Scalar Regression

*Zikai Lin, ♦Yajuan Si and Jian Kang.* University of Michigan

15:50 Mediation Analysis for High-Dimensional Mediators and Outcomes with an Application to Multimodal Imaging Data

*Zhiwei Zhao and ♦Shuo Chen.* University of Maryland

16:15 Statistical Inferences for Complex Dependence of Multi-modal Imaging Data

*Jinyuan Chang<sup>1</sup>, ♦Jing He<sup>1</sup>, Jian Kang<sup>2</sup> and Mingcong Wu<sup>1</sup>.* <sup>1</sup>Southwestern University of Finance and Economics <sup>2</sup>University of Michigan

16:40 Floor Discussion.

**Session 23CHI7: Data Borrowing : Methodology and Application**

Room: SONGJIANG HALL

Organizer: Qian Jiang, Department of Biostatistics and Programming, China, Sanofi, Inc..

Chair: Qian Jiang, Department of Biostatistics and Programming, China, Sanofi, Inc..

15:00 Bayesian Borrowing from Historical Control Data in Vaccine Efficacy Trial

*Penny Peng.* Department of Biostatistics and Programming, China, Sanofi, Inc.

- 15:25 A Brief Introduction of Historical Data Borrowing Approach Implemented in a Real Case Study.  
♦*Gaowei Nian, Ning Li and Genming Shi.*
- 15:50 Methods of Reconstructing Individual Patients Data and Subgroup Survival Curves from Published Kaplan-Meier Plots  
*Sheng Xu.* BeiGene
- 16:15 Reducing Bias using Propensity Score-Integrated Composite Likelihood Approach for Incorporating Multiple External Data Sources  
♦*Leixin Xia<sup>1</sup>, Yishen Yang<sup>2</sup>, Weilong Zhao<sup>1</sup>, Chaohui Yuan<sup>1</sup>, Ming Chen<sup>1</sup>, Wei Tan<sup>3</sup>, Zhini Wang<sup>2</sup> and Bin Jia<sup>1</sup>.* <sup>1</sup>Janssen <sup>2</sup>IQVIA <sup>3</sup>ICON plc
- 16:40 Floor Discussion.

### Session 23CHI73: Recent Developments in Medical Bioinformatics

Room: BAOSHAN HALL

Organizer: Yichuan Zhao, Department of Mathematics & Statistics, Georgia State University, Atlanta, GA, USA.

Chair: Yucong Lin, Institute of Engineering Medicine, Beijing Institute of Technology, Beijing, China.

- 15:00 Biomedical Entity Linking by Text Generation and Knowledge Enhancement

♦*Hongyi Yuan, Zheng Yuan and Sheng Yu.* Tsinghua University

- 15:25 Multimodal Learning on Graphs for Disease Relation Extraction

♦*Yucong Lin<sup>1</sup>, Keming Lu<sup>2</sup>, Sheng Yu<sup>3</sup>, Tianxi Cai<sup>4</sup> and Marinka Zitnik<sup>4</sup>.* <sup>1</sup>Institute of Engineering Medicine, Beijing Institute of Technology, Beijing, China <sup>2</sup>cViterbi School of Engineering, University of Southern California, Los Angeles, USA <sup>3</sup>eDepartment of Industrial Engineering, Tsinghua University, Beijing, China <sup>4</sup>gDepartment of Biomedical Informatics, Harvard Medical School, Boston, MA, 02115, USA

- 15:50 Knowledge Graph Embedding with Electronic Health Records Data

♦*Junwei Lu, Tianxi Cai and Doudou Zhou.* Harvard T.H. Chan School of Public Health

- 16:15 Rwe-Ready: Pipeline to Harness Electronic Health Records for Real-World Evidence

♦*Jue Hou<sup>1</sup>, Rachel Zhao<sup>2</sup>, Jessica Gronsbell<sup>3</sup>, Yucong Lin<sup>4</sup>, Clara-Lea Bonzel<sup>5</sup>, Qingyi Zeng<sup>1</sup>, Sinian Zhang<sup>6</sup>, Brett Beaulieu-Jones<sup>7</sup>, Griffin Webber<sup>5</sup> and And Others<sup>8</sup>.*

<sup>1</sup>University of Minnesota <sup>2</sup>University of British Columbia

<sup>3</sup>University of Toronto <sup>4</sup>Beijing Institute of Technology

<sup>5</sup>Harvard Medical School <sup>6</sup>Renmin University of China

<sup>7</sup>University of Chicago <sup>8</sup>Merck & Co, Duke University, Harvard Medical School

- 16:40 Floor Discussion.

### Session 23CHI77: Recent Advances in Biostatistics

Room: XUHUI HALL

Organizer: Haitao Zheng, Southwest Jiaotong University.

Chair: Lei Huang, Southwest Jiaotong University.

- 15:00 Sc-Meb: Spatial Clustering with Hidden Markov Random Field using Empirical Bayes

♦*Yang Yi<sup>1</sup>, Xingjie Shi<sup>2</sup>, Wei Liu<sup>1</sup>, Qiuzhong Zhou<sup>3</sup>, Mai Chan Lau<sup>4</sup>, Jeffrey Chun Tatt Lim<sup>4</sup>, Lei Sun<sup>5</sup>, Cedric Chuan Young Ng<sup>6</sup>, Joe Yeong<sup>7</sup> and Liu Liu<sup>8</sup>.* <sup>1</sup>Health Services & Systems Research program at Duke-NUS Medical School <sup>2</sup>Academy of Statistics and Interdisciplinary Sciences at East China Normal University, Shanghai, China <sup>3</sup>Cardiovascular and Metabolic Disorders program at Duke-NUS Medical School, Singapore. <sup>4</sup>IMCB, ASTAR Singapore. <sup>5</sup>Cardiovascular and Metabolic Disorders program at Duke-NUS Medical School, Singapore. <sup>6</sup>CTO of Cancer Discovery Hub, National Cancer Centre Singapore <sup>7</sup>IMCB, ASTAR Singapore <sup>8</sup>Health Services & Systems Research program at Duke-NUS Medical School, Singapore.

- 15:25 Sufficient Variable Screening for Ultrahigh Dimensional Right Censored Data via Independence Measures

♦*Baoying Yang<sup>1</sup>, Qingcong Yuan and Xiangrong Yin.* <sup>1</sup>Department of Statistics, College of Mathematics, Southwest Jiaotong University

- 15:50 Mendelian Randomization Accounting for Complex Correlated Horizontal Pleiotropy While Elucidating Shared Genetic Etiology

♦*Qing Cheng<sup>1</sup>, Xiao Zhang, Lin Chen<sup>2</sup> and Jin Liu<sup>3</sup>.* <sup>1</sup>Center of Statistical Research, School of Statistics, Southwestern University of Finance and Economics, Chengdu, Sichuan, China. <sup>2</sup>Department of Public Health Sciences, The University of Chicago, Chicago, IL, USA. <sup>3</sup>Centre for Quantitative Medicine, Health Services & Systems Research, Duke-NUS Medical School, Singapore, Singapore

- 16:15 A General Framework for Identifying Hierarchical Interactions and Its Application to Genomics Data

♦*Xiao Zhang<sup>1</sup>, Xingjie Shi<sup>2</sup>, Yiming Liu<sup>3</sup>, Xu Liu<sup>3</sup> and Shuangge Ma<sup>4</sup>.* <sup>1</sup>School of Data Science, The Chinese University of Hong Kong-Shenzhen <sup>2</sup>KLATASDS-MOE, Academy of Statistics and Interdisciplinary Sciences, East China Normal University <sup>3</sup>School of Statistics and Management, Shanghai University of Finance and Economics <sup>4</sup>Department of Biostatistics, Yale University

- 16:40 Discussant: Yi Yang, National University of Singapore, Duke-NUS Medical School

Floor Discussion.

### Session 23CHI84: Advances in Causal Machine Learning: Challenges and Breakthrough

Room: LONGFENG HALL-III

Organizer: Hengrui Cai, Department of Statistics, University of California Irvine.

Chair: Hengrui Cai, Department of Statistics, University of California Irvine.

- 15:00 Matrix-Valued Network Autoregression Model with Latent Group Structure

Yimeng Ren<sup>1</sup>, ♦Xuening Zhu<sup>1</sup>, Ganggang Xu<sup>2</sup> and Yanyuan Ma<sup>3</sup>. <sup>1</sup>Fudan University <sup>2</sup>University of Miami <sup>3</sup>The Pennsylvania State University

15:25 Sparse Causal Mediation Analysis with Unmeasured Mediator-Outcome Confounding  
*Kang Shuai<sup>1</sup>, Lan Liu<sup>2</sup>, Yangbo He<sup>1</sup> and ♦Wei Li<sup>3</sup>. <sup>1</sup>Peking University <sup>2</sup>University of Minnesota <sup>3</sup>Renmin University of China*

15:50 Optimal Individualized Decision-Making with Proxies  
♦Tao Shen<sup>1</sup> and Yifan Cui<sup>2</sup>. <sup>1</sup>Department of Statistics and Data Science, National University of Singapore <sup>2</sup>Center for Data Science, Zhejiang University

16:15 Blessing from Human-AI Interaction: Super Reinforcement Learning in Confounded Environments  
♦Jiayi Wang<sup>1</sup>, Zhengling Qi<sup>2</sup> and Chengchun Shi<sup>3</sup>. <sup>1</sup>University of Texas at Dallas <sup>2</sup>George Washington University <sup>3</sup>London School of Economics and Political Science

16:40 Floor Discussion.

### Session 23CHI85: Large-Scale Dependent Data Modelling and Analysis

Room: JINSHAN HALL

Organizer: Xuening Zhu, Fudan University.

Chair: Xuening Zhu, Fudan University.

15:00 Sequential One-Step Estimator by Subsampling for Customer Churn Analysis with Massive Datasets  
♦Feifei Wang<sup>1</sup>, Danyang Huang<sup>1</sup>, Tianchen Gao<sup>2</sup>, Shuyuan Wu<sup>3</sup> and Hansheng Wang<sup>3</sup>. <sup>1</sup>Renmin University of China <sup>2</sup>Xiamen University <sup>3</sup>Peking University

15:25 Transfer Learning for Spatial Autoregressive Models  
Hao Zeng, ♦Wei Zhong and Xingbai Xu. Xiamen University

15:50 Adaptive False Discovery Rate Control with Privacy Guarantee  
Xintao Xia<sup>1</sup> and ♦Zhanrui Cai<sup>2</sup>. <sup>1</sup>Iowa State University <sup>2</sup>University of Hong Kong

16:15 Distributed Logistic Regression for Massive Data with Rare Events  
♦Xuetong Li<sup>1</sup>, Xuening Zhu<sup>2</sup> and Hansheng Wang<sup>1</sup>. <sup>1</sup>Peking University <sup>2</sup>Fudan University

16:40 Floor Discussion.

### Session 23CHI86: Modern Statistics on Complex Data

Room: PUTUO HALL

Organizer: Yiyuan She, Florida State University.

Chair: Yiyuan She, Florida State University.

15:00 Nsr-Ucd Model for Covid-19 Transmission with Unaware Infections.  
♦Chang Liu<sup>1</sup> and Yiyuan She<sup>2</sup>. <sup>1</sup>Jilin University <sup>2</sup>Florida State University

15:25 Bayesian Hierarchical Model for Patient-Specific Abnormal Region Detection  
Rongjie Liu. Florida State University

15:50 Pivot Statistics for Normal Populations  
Liqiang Ni. University of Central Florida

16:15 Consistent Dynamic Bayesian Network Learning for a Fmri Study of Emotion Processing  
♦Lizhe Sun<sup>1</sup>, Aiying Zhang<sup>2</sup> and Faming Liang<sup>3</sup>. <sup>1</sup>Peking University <sup>2</sup>Columbia University and New York State Psychiatric Institute <sup>3</sup>Purdue University

16:40 A Sparse Ising Model with Latent Variables  
Lizhu Tao.  
Floor Discussion.

**July 1, 17:00-18:40**

### Session 23CHI101: Advances in Analysis of Genomic and High Dimensional Data Analysis

Room: NANSHI HALL

Organizer: Heping Zhang, Yale University.

Chair: Heping Zhang, Yale University.

17:00 Likelihood Ratio Test for Poisson Directed Acyclic Graph  
Shuyan Chen<sup>1</sup>, Xin Liu<sup>2</sup>, Xiaotong Shen<sup>3</sup> and ♦Shaoli Wang<sup>2</sup>. <sup>1</sup>University of Science and Technology of China <sup>2</sup>Shanghai University of Finance and Economics <sup>3</sup>University of Minnesota

17:25 Development of Pulmonary Nodule Malignancy Risk Models using Nlst Data  
Fenghai Duan. Brown University School of Public Health

17:50 A Statistical Learning Method for Simultaneous Copy Number Estimation and Subclone Clustering with Single Cell Sequencing Data  
Fei Qin<sup>1</sup>, Guoshuai Cai<sup>1</sup> and ♦Feifei Xiao<sup>2</sup>. <sup>1</sup>University of South Carolina <sup>2</sup>University of Florida

18:15 Dna Methylation in the Eyes of a Biological Data Scientist: From Biomarkers to Functional Interpretation  
Xiang Chen. St. Jude Children's Research Hospital

18:40 Floor Discussion.

### Session 23CHI102: Frontiers in -Omics Data Analysis

Room: PUTUO HALL

Organizer: Chong Jin, New Jersey Institute of Technology.

Chair: Tian Tian, Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA.

17:00 Analysis of Spatial Transcriptomics using Optimal Transport  
Zixuan Cang. North Carolina State University

17:25 Spatially Aware Dimension Reduction for Spatial Transcriptomics  
♦Lulu Shang and Xiang Zhou. University of Michigan

17:50 Spatial Dependency-Aware Deep Generative Models  
♦Tian Tian<sup>1</sup>, Jie Zhang<sup>2</sup>, Xiang Lin<sup>2</sup>, Zhi Wei<sup>2</sup> and Hakon Hakonarson<sup>1</sup>. <sup>1</sup>Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA. <sup>2</sup>Department of Computer Science, New Jersey Institute of Technology, Newark, New Jersey, USA.

18:15 Genome-Wide Prediction of Structural Variations and Enhancer-Hijacking Events from Chromatin Interaction Data in Cancer Genomes

<sup>♦</sup>Xiaotao Wang<sup>1</sup>, Jie Xu<sup>2</sup>, Baozhen Zhang<sup>3</sup>, Yu Luan<sup>4</sup>, Fan Song<sup>5</sup>, Hongbo Yang<sup>1</sup>, Juan Wang<sup>4</sup>, Tingting Liu<sup>6</sup> and Feng Yue<sup>4</sup>. <sup>1</sup>Fudan University <sup>2</sup>University of California San Diego <sup>3</sup>Peking University Cancer Hospital and Institute <sup>4</sup>Northwestern University <sup>5</sup>Illumina <sup>6</sup>Southeast University

18:40 Floor Discussion.

### Session 23CHI156: Mitigating Incomplete Data Biases-a Modern Take

Room: JIADING HALL

Organizer: Shu Yang, North Carolina State University.

Chair: Wang Miao, Peking University.

17:00 Robust and Efficient Case-Control Studies with Contaminated Case Pools: a Unified m-Estimation Framework

<sup>♦</sup>Guorong Dai<sup>1</sup> and Jinbo Chen<sup>2</sup>. <sup>1</sup>Fudan University <sup>2</sup>University of Pennsylvania

17:25 Floor Discussion.

### Session 23CHI163: New Generation of Statisticians in Drug Development

Room: NANHUI HALL

Organizer: Gang Li, Eisai Inc, USA.

Chair: James Pan, Janssen(China) R& D, China.

17:00 Panelists: Dr. Michael Lee, Harbour Biomed, China;

Dr. Chen Ji, Sanofi, China;

Dr. Yiming Chen, Janssen (China) R&D, China

### Session 23CHI167: Causal Inference: From Practice to Theory

Room: FENGXIAN HALL

Organizer: Peng Ding, University of California Berkeley.

Chair: Peng Ding, University of California Berkeley.

17:00 Inference of Treatment Effects and Complier Treatment Effects under Computer Assisted Balance-Improving Designs

<sup>♦</sup>Junni Zhang<sup>1</sup>, Per Johansson<sup>2</sup> and Zhen Zhong<sup>3</sup>.

<sup>1</sup>National School of Development, Peking University

<sup>2</sup>Department of Statistics, Uppsala University <sup>3</sup>Department of Mathematical Sciences, Tsinghua University

17:25 Interpretable Sensitivity Analysis for the Baron-kenny Approach to Mediation with Unmeasured Confounding

Mingrui Zhang and <sup>♦</sup>Peng Ding. University of California Berkeley

17:50 Optimal Treatment Regime under the Individual No-Harm Criterion

Peng Wu. Beijing Technology and Business University

18:15 Floor Discussion.

### Session 23CHI174: Special Memorial Session to Celebrate Life of Professor Tze Leung Lai

Room: YANGPU HALL

Organizer: Ying Lu, Stanford University; Zhiliang Ying, Columbia University; Dylan Small, University of Pennsylvania; Zhezhen Jin, Columbia University.

Chair: Zhezhen Jin, Columbia University.

17:00 深切怀念黎先生

Dong Han. Shanghai Jiaotong University

17:20 Causal IV Analysis for Time-to-Event Data

Gang Li. UCLA

17:40 Moment Deviation Subspaces of Dimension Reduction for High-Dimensional Data with Change Structure

<sup>♦</sup>Xuehu Zhu<sup>1</sup>, Luoyao Yu<sup>1</sup>, Jiaqi Huang<sup>2</sup>, Junmin Liu<sup>1</sup> and Lixing Zhu<sup>2</sup>. <sup>1</sup>Xi'an Jiaotong University <sup>2</sup>Beijing Normal University

18:10 Best Arm Identification in Batched Multi-Armed Bandit Problems

Shengyu Cao, Simai He, Jin Xu and <sup>♦</sup>Hongsong Yuan. Shanghai University of Finance and Economics

18:30 Panelists: Dong Han, Shanghai Jiao Tong University, Shanghai, China;

Gang Li, University of California, Los Angeles, USA;

Hongsong Yuan, Shanghai University of Finance and Economics, Shanghai, China;

Ka Wai Tsang, Chinese University of Hong Kong (Shenzhen), China

### Session 23CHI158: From Early Development to PAC: Application of Bayesian Analysis in Drug Development

Room: CHANGNING HALL

Organizer: Hongjie Deng, Boehringer Ingelheim.

Chair: Huiyao Huang, Cancer Hospital Chinese Academy of Medical Sciences.

17:00 A Bayesian Sample Size Planning Tool for Phase i Dose-Finding Trials

Xiaolei Lin. Fudan University

17:25 Application of Bayesian Analysis in Early Phase Drug Development

Chengyuan Song. boehringer-ingelheim

17:50 Basic: a Bayesian Adaptive Synthetic Control Design for Phase II Clinical Trials

<sup>♦</sup>Liyun Jiang, Peter F Thall, Fangrong Yan, Scott Kopetz and Ying Yuan.

18:15 Bayesian Dynamic Borrowing Method Implementation in a Phase III Post-Marketing Trial

Ling Li.

18:40 Floor Discussion.

### Session 23CHI160: Novel Machine Learning Methods to Advance Precision Medicine using Big Biomarker Data

Room: LONGFENG HALL-II

Organizer: Shanghong Xie, Southwestern University of Finance and Economics.

Chair: Shiyang Ma, Shanghai Jiao Tong University.

17:00 A Flexible Summary-Based Colocalization Method with Application to the Mucin Cystic Fibrosis Lung Disease Modifier Locus

<sup>♦</sup>Fan Wang<sup>1</sup>, Naim Panjwani<sup>2</sup>, Cheng Wang<sup>2</sup>, Lei Sun<sup>3</sup> and Lisa Strug<sup>3</sup>. <sup>1</sup>Columbia University <sup>2</sup>The Hospital for Sick Children <sup>3</sup>University of Toronto

- 17:25 Blockwise Mixed Membership Models for Multivariate Longitudinal Categorical Data  
 ♦*Kai Kang*<sup>1</sup> and *Yuqi Gu*<sup>2</sup>. <sup>1</sup>Sun Yat-sen University  
<sup>2</sup>Columbia University
- 17:50 Knockoff-Based Statistics for the Identification of Putative Causal Genes in Genetic Studies  
 ♦*Shiyang Ma*<sup>1</sup>, *Linxi Liu*<sup>2</sup>, *Zihuai He*<sup>3</sup> and *Iuliana Ionita-Laza*<sup>4</sup>. <sup>1</sup>Shanghai Jiao Tong University School of Medicine  
<sup>2</sup>University of Pittsburgh <sup>3</sup>Stanford University <sup>4</sup>Columbia University
- 18:15 Longitudinal Classification and Forecast in the Absence of True Disease Labels  
 ♦*Zexi Cai*<sup>1</sup>, *Karen Marder*<sup>2</sup>, *Donglin Zeng*<sup>3</sup> and *Yuanjia Wang*<sup>1</sup>. <sup>1</sup>Department of Biostatistics, Columbia University <sup>2</sup>Department of Neurology, Columbia University <sup>3</sup>Department of Biostatistics, University of North Carolina at Chapel Hill
- 18:40 Floor Discussion.

### Session 23CHI63: Technical Advances on Analyzing Single Cell RNA-Seq and Spatial Transcriptomic Data

Room: T-1

Organizer: Zhenxing Guo, School of Data Science, Chinese University of Hong Kong, Shenzhen.  
 Chair: Zhenxing Guo, School of Data Science, Chinese University of Hong Kong, Shenzhen.

- 17:00 Intelligent Spatial Transcriptomics: Paving the Way for Deciphering Tissue Architecture  
*Shihua Zhang*.
- 17:25 Identifying Phenotype-Associated Subpopulations by Integrating Bulk and Single-Cell Sequencing Data  
*Duanchen Sun*.
- 17:50 Integrative Models of Single Cell Genomics Data for Dissecting Cellular Heterogeneity and Transcriptional Regulation  
*Lihua Zhang*. Wuhan University
- 18:15 Spider: a Flexible and Unified Framework for Simulating Spatial Transcriptomics Data  
*Xiaoqi Zheng*. Shanghai Jiao Tong University School of Medicine
- 18:40 Floor Discussion.

### Session 23CHI64: Model Estimation and Hypothesis Testing of High Dimensional Data

Room: XUHUI HALL

Organizer: Jing Yang, Hunan Normal University.  
 Chair: Jing Yang, Hunan Normal University.

- 17:00 Use of Random Integration to Test Equality of High Dimensional Covariance Matrices  
 ♦*Jiang Yunlu*<sup>1</sup>, *Wen Canhong*<sup>2</sup>, *Jiang Yukang*<sup>3</sup>, *Wang Xueqin*<sup>2</sup> and *Zhang Heping*<sup>4</sup>. <sup>1</sup>Jinan University <sup>2</sup>University of Science and Technology of China <sup>3</sup>Sun Yat-Sen University  
<sup>4</sup>Yale University

- 17:25 Robust Optimal Subsampling Based on Weighted Asymmetric Least Squares  
*Min Ren*<sup>1</sup>, *Shengli Zhao*<sup>1</sup>, ♦*Mingqiu Wang*<sup>1</sup> and *Xinbei Zhu*<sup>2</sup>. <sup>1</sup>Qufu Normal University <sup>2</sup>Virginia Tech University
- 17:50 A Scale-Invariant Test on General Linear Hypothesis in High-Dimensional Heteroscedastic One-Way Manova  
 ♦*Mingxiang Cao*<sup>1</sup> and *Ziyang Cheng*<sup>2</sup>. <sup>1</sup>Anhui Normal University <sup>2</sup>Changchun University of Technology
- 18:15 Relative Error Model Average  
 ♦*Xiaochao Xia*<sup>1</sup>, *Hao Ming*<sup>1</sup> and *Jialiang Li*<sup>2</sup>. <sup>1</sup>Chongqing University <sup>2</sup>National University of Singapore
- 18:40 Floor Discussion.

### Session 23CHI87: Advances in Machine Learning with Applications

Room: LONGFENG HALL-III

Organizer: Yiyuan She, Florida State University.  
 Chair: Yiyuan She, Florida State University.

- 17:00 Properties of Standard and Sketched Kernel Fisher Discriminant  
*Heng Lian*.
- 17:25 Skewed Pivotal-Point-Adaptive Modeling with Applications to Semicontinuous Outcomes  
 ♦*Yiyuan She*, *Xiaoqiang Wu*, *Lizhu Tao* and *Debajyoti Sinha*.
- 17:50 when Mediation Analysis Faces Subgroup Heterogeneity  
*Kaizhou Lei*<sup>1</sup>, *Shengxian Ding*<sup>1</sup>, *Lexin Li*<sup>2</sup>, *Rongjie Liu*<sup>1</sup> and ♦*Chao Huang*<sup>1</sup>. <sup>1</sup>Florida State University <sup>2</sup>UC Berkley
- 18:15 Single Index Models with Regularized Matrix Coefficients  
*Luo Xiao*. North Carolina State University
- 18:40 Floor Discussion.

### Session 23CHI90: Advances in Statistical Methodologies for Clustered Data Analysis and Clustered Randomized Trials Design

Room: LONGFENG HALL-I

Organizer: Dongdong Li, Department of Population Medicine, Harvard Medical School, Boston, MA USA..  
 Chair: Dongdong Li, Department of Population Medicine, Harvard Medical School, Boston, MA USA..

- 17:00 Estimation and Inference for Complexly Correlated Data via Network Generalized Estimating Equations  
 ♦*Tom Chen*<sup>1</sup>, *Fan Li*<sup>2</sup> and *Rui Wang*<sup>1</sup>. <sup>1</sup>Harvard Medical School <sup>2</sup>Yale University
- 17:25 Marginal Proportional Hazards Models for Clustered Interval-Censored Data with Time-Dependent Covariates  
 ♦*Kaitlyn Cook*<sup>1</sup>, *Wenbin Lu*<sup>2</sup> and *Rui Wang*<sup>3</sup>. <sup>1</sup>Smith College <sup>2</sup>North Carolina State University <sup>3</sup>Harvard University
- 17:50 Marginal Structural Models for Network-Level Interventions: The Limiting Variance is Smaller with Estimated Weights  
 ♦*Judith Lok*<sup>1</sup>, *Ashley Buchanan*<sup>2</sup>, *Luis Iberico*<sup>1</sup> and *Donna Spiegelman*<sup>3</sup>. <sup>1</sup>Boston University <sup>2</sup>University of Rhode Island <sup>3</sup>Yale University

18:15 Model-Robust and Efficient Covariate Adjustment for Cluster-Randomized Experiments

*Bingkai Wang<sup>1</sup>, Chan Park<sup>1</sup>, Dylan Small<sup>1</sup> and ♦Fan Li<sup>2</sup>.*

<sup>1</sup>The Statistics and Data Science Department of the Wharton School, University of Pennsylvania, Philadelphia, PA, USA

<sup>2</sup>Department of Biostatistics, Yale University School of Public Health, New Haven, CT, USA

18:40 Floor Discussion.

### Session 23CHI91: Recent Development in Estimating Treatment Effects with Complex Medical Data

Room: JINGAN HALL

Organizer: Tom Chen, Department of Population Medicine, Harvard Medical School, Boston, MA USA..

Chair: Tom Chen, Department of Population Medicine, Harvard Medical School, Boston, MA USA..

17:00 Examining Subgroup-Specific Treatment Effects in Multi-Source Data: Source-Specific Inference and Transportability to an External Population

♦*Guanbo Wang<sup>1</sup>, Alex Levis<sup>2</sup> and Issa Dahabreh<sup>1</sup>.*

<sup>1</sup>Harvard University <sup>2</sup>Carnegie Mellon University

17:25 Modeling Interval-Censored Outcome Data with an Interval-Censored Covariate with Application to Hiv Viral Bound Research

♦*Dongdong Li<sup>1</sup>, Yue Song<sup>2</sup>, Wenbin Lu<sup>3</sup>, Huldrych Gunthard<sup>4</sup>, Roger Kouyos<sup>4</sup> and Rui Wang<sup>5</sup>.* <sup>1</sup>Harvard Medical School <sup>2</sup>Harvard T. H. Chan School of Public Health <sup>3</sup>North Carolina State University <sup>4</sup>University of Zurich <sup>5</sup>Harvard Medical School and Harvard T. H. Chan School of Public Health

17:50 Causal Inference for Assessing Cost-Effectiveness with Multi-State Modelling

♦*Yi Xiong<sup>1</sup>, Gary Chan<sup>2</sup> and Li Hsu<sup>3</sup>.* <sup>1</sup>University of Manitoba; Fred Hutchinson Cancer Center <sup>2</sup>University of Washington <sup>3</sup>Fred Hutchinson Cancer Center

18:15 Floor Discussion.

### Session 23CHI92: Advanced Methods for Analyzing Categorical Data

Room: LUWAN HALL

Organizer: Juxin Liu, University of Saskatchewan.

Chair: Juxin Liu, University of Saskatchewan.

17:00 Inference for Seemingly Unrelated Linear Mixed Models

♦*Lichun Wang and Yang Yang.* Department of Statistics, Beijing Jiaotong University

17:25 Classification with Imbalanced Data in Medicine

*Wanhua Su.* MacEwan University

17:50 Iteratively Reweighted Least Squares Method for Estimating Polyserial and Polychoric Correlation Coefficients

*Peng Zhang<sup>1</sup>, ♦Ben Liu<sup>1</sup> and Jingjing Pan<sup>2</sup>.* <sup>1</sup>School of Mathematical Sciences, Zhejiang University <sup>2</sup>Zhejiang Super Soul Artificial Intelligence Research Institute

18:15 Modeling Clustered Binary Data with an Application to Stock Crash Analysis

*Ruixi Zhao<sup>1</sup>, Renjun Ma<sup>1</sup>, ♦Guohua Yan<sup>1</sup>, Haomiao Niu<sup>2</sup> and Wenjiang Jiang<sup>3</sup>.* <sup>1</sup>University of New Brunswick <sup>2</sup>University of Siena <sup>3</sup>Yunan Normal University

18:40 Floor Discussion.

### Session 23CHI93: Limit Theorems and Inference for Time Series and Spatial Processes

Room: SONGJIANG HALL

Organizer: Yimin Xiao, Michigan State University.

Chair: Yimin Xiao, Michigan State University.

17:00 Limit Theory for Autoregressive Processes with Roots Close to Unity

*Nannan Ma<sup>1</sup>, Hailin Sang<sup>2</sup> and ♦Guangyu Yang<sup>3</sup>.*

<sup>1</sup>Agricultural Bank of China <sup>2</sup>University of Mississippi

<sup>3</sup>Zhengzhou University

17:25 Local Limit Theorem for Linear Random Fields

*Timothy Fortune<sup>1</sup>, Magda Peligrad<sup>2</sup>, ♦Hailin Sang<sup>3</sup>, Yimin Xiao<sup>4</sup> and Guangyu Yang<sup>5</sup>.* <sup>1</sup>Nicholls State University <sup>2</sup>University of Cincinnati <sup>3</sup>University of Mississippi

<sup>4</sup>Michigan State University <sup>5</sup>Zhengzhou University

17:50 Kernel Entropy Estimation for Long Memory Linear Processes with Infinite Variance

*Hui Liu and ♦Fangjun Xu.* East China Normal University

18:15 Self-Normalized Cramer Type Moderate Deviations for Martingales with Applications

♦*Xiequan Fan<sup>1</sup> and Qiman Shao<sup>2</sup>.* <sup>1</sup>Northeastern University at Qinhuangdao <sup>2</sup>Southern University of Science and Technology

18:40 Floor Discussion.

### Session 23CHI96: New Advances in Causal Inference and Missing Data

Room: JINSHAN HALL

Organizer: Liangyuan Hu, Rutgers University.

Chair: Liangyuan Hu, Rutgers University.

17:00 Bayesian Machine Learning g-Computation Formula for Survival Data

♦*Xinyuan Chen<sup>1</sup>, Liangyuan Hu<sup>2</sup> and Fan Li<sup>3</sup>.* <sup>1</sup>Mississippi State University <sup>2</sup>Rutgers University <sup>3</sup>Yale University

17:25 Estimation of Causal Influence Effect in Social Networks with Latent Location Adjustment

♦*Samrachana Adhikari and Seungha Um.* NYU School of Medicine

17:50 Estimating the Causal Effect of a Longitudinal Treatment when Covariates are Subject to Missing Data

♦*Liangyuan Hu<sup>1</sup> and Jungang Zou<sup>2</sup>.* <sup>1</sup>Rutgers University <sup>2</sup>Columbia University

18:15 Floor Discussion.

### Session 23CHI98: Analysis of Complex Network and Text Data

Room: BAOSHAN HALL

Organizer: Jiashun Jin, Carnegie Mellon University.

Chair: Jiashun Jin, Carnegie Mellon University.

17:00 Bayesian Inference of Spatially Varying Correlations via the Thresholded Correlation Gaussian Process

♦*Moyan Li<sup>1</sup>, Jian Kang<sup>1</sup> and Lexin Li<sup>2</sup>.* <sup>1</sup>University of Michigan <sup>2</sup>University of Berkley

17:25 Stock Co-Jump Networks

\*Yi Ding<sup>1</sup>, Guoli Li<sup>2</sup>, Yingying Li<sup>2</sup> and Xinghua Zheng<sup>2</sup>.

<sup>1</sup>University of Macau <sup>2</sup>Hong Kong University of Science and Technology

17:50 Floor Discussion.

### Session 23CHI99: Advances in Bayesian Adaptive Design Methods for Drug Development

Room: CHONGMING HALL

Organizer: Rongji Mu, Shanghai Jiao Tong University.

Chair: Hongjie Deng, Boehringer-Ingelheim.

17:00 From Finding Maximum Tolerated Dose (Mtd) to Determining Optimal Biological Dose (Obd) in Oncology Drug Development

J. Jack Lee. University of Texas MD Anderson Cancer Center

17:25 Demo: Bayesian Adaptive Dose Exploration – monitoring–optimization Design Based on Short, Intermediate, and Long-Term Outcomes

Ruitao Lin. The University of Texas MD Anderson Cancer Center

17:50 A Bayesian Adaptive Phase i/Ii Clinical Trial Design with Late-Onset Competing Risk Outcomes

\*Yifei Zhang<sup>1</sup>, Sha Cao<sup>2</sup>, Chi Zhang<sup>2</sup>, Ick Hoon Jin<sup>3</sup> and Yong Zang<sup>2</sup>. <sup>1</sup>Jiangsu Hengrui Pharmaceuticals Co., Ltd. <sup>2</sup>Indiana University, Biostatistics & Health Data Science <sup>3</sup>Yonsei University

18:15 Shotgun-2: a Bayesian Phase i/Ii Basket Trial Design to Identify Indication-Specific Optimal Biological Doses

Fangrong Yan.

18:40 Floor Discussion.

### July 2, 8:30-9:30

### Session 23CHIKT2: Keynote Lecture 2

Room: LONGFENG HALL

Organizer: ICSA China Conference Organizing Committee.

Chair: Zhezhen Jin, Columbia University.

8:30 Theory of Fpca for Discretized Functional Data

Fang Yao. Peking University

9:20 Floor Discussion.

### July 2, 10:00-11:40

### Session 23CHI123: Recent Developments in Causal Inference

Room: SONGJIANG HALL

Organizer: Xingdong Feng, Shanghai University of Finance and Economics, China.

Chair: Wendong Li, Shanghai University of Finance and Economics, China.

10:00 Estimation and Inference of a Directed Acyclic Graph using Gwas Summary Statistics

Rachel Zilinskas<sup>1</sup>, \*Chunlin Li<sup>2</sup>, Xiaotong Shen<sup>2</sup>, Wei Pan<sup>3</sup> and Tianzhong Yang<sup>3</sup>. <sup>1</sup>Statistics and Data Corporation

<sup>2</sup>School of Statistics, University of Minnesota <sup>3</sup>Division of Biostatistics, University of Minnesota

10:25 Local Method for Causal Effect Estimation

Yue Liu.

10:50 Learning Linear Non-Gaussian Directed Acyclic Graph with Diverging Number of Nodes

Ruixuan Zhao, \*Xin He and Junhui Wang.

11:15 Introducing the Specificity Score: a Measure of Causality Beyond p Value

Wang Miao. Peking University

11:40 Discussant: Wang Miao, Peking University

Floor Discussion.

### Session 23CHI130: Space-Filling Designs (I)

Room: JINGAN HALL

Organizer: Jian-Feng Yang, Nankai University.

Chair: Jian-Feng Yang, Nankai University.

10:00 Column Expanded Latin Hypercube Designs

\*Qiao Wei, Jian-Feng Yang and Min-Qian Liu. Nankai University

10:25 Construction of Orthogonal Doubly Coupled Designs

Mengmeng Liu, \*Jinyu Yang and Min-Qian Liu. Nankai University

10:50 A-Optimal Designs for Non-Parametric Symmetrical Global Sensitivity Analysis

\*Xueping Chen<sup>1</sup>, Yujie Gai<sup>2</sup> and Xiaodi Wang<sup>2</sup>. <sup>1</sup>Jiangsu University of Technology <sup>2</sup>Central University of Finance and Economics

11:15 Construction of a Class of Nested Orthogonal Arrays

Shanqi Pang. Henan Normal University

11:40 Floor Discussion.

### Session 23CHI135: Tree and Graphical Models

Room: NANHUI HALL

Organizer: Annie Qu, UC Irvine.

Chair: Junhui Wang, Chinese University of Hong Kong.

10:00 Chain Graph Models: Identifiability, Estimation and Asymptotics

Ruixuan Zhao<sup>1</sup>, Haoran Zhang<sup>2</sup> and \*Junhui Wang<sup>2</sup>. <sup>1</sup>City University of Hong Kong <sup>2</sup>Chinese University of Hong Kong

10:25 Confidence Band Estimation of Random Survival Forests

Sarah Formentini<sup>1</sup>, Wei Liang<sup>2</sup> and \*Ruoqing Zhu<sup>1</sup>. <sup>1</sup>University of Illinois Urbana Champaign <sup>2</sup>Xiamen University

10:50 Joint Modeling of Change-Point Identification and Dependent Dynamic Community Detection

\*Diqing Li<sup>1</sup>, Yubai Yuan<sup>2</sup>, Xinsheng Zhang<sup>3</sup> and Annie Qu<sup>4</sup>. <sup>1</sup>Zhejiang Gongshang University <sup>2</sup>the Pennsylvania State University <sup>3</sup>Fudan University <sup>4</sup>University of California Irvine

- 11:15 Tree Building via Hypothesis Tests  
*Ze Gao<sup>1</sup>, Bo Zhang<sup>2</sup>, Jiaqi Hu<sup>2</sup>, Tingyin Wang<sup>2</sup>, Heping Zhang<sup>3</sup> and ♦Xueqin Wang<sup>2</sup>. <sup>1</sup>Sun Yat-sen University <sup>2</sup>University of Science and Technology of China <sup>3</sup>Yale University*
- 11:40 Floor Discussion.

### Session 23CHI161: Modelling Unstructured Data: Image, Network, Text and Beyond

Room: T-1

Organizer: Junhui Wang, The Chinese University of Hong Kong.  
 Chair: Ben Dai, The Chinese University of Hong Kong.

- 10:00 Dimension Reduction for Covariates in Network Data  
*♦Junlong Zhao<sup>1</sup>, Xumin Liu<sup>2</sup>, Hansheng Wang<sup>3</sup> and Chenlei Leng<sup>4</sup>. <sup>1</sup>Beijing Normal University <sup>2</sup>Beijing Technology and Business University <sup>3</sup>Peking University <sup>4</sup>University of Warwick*
- 10:25 Rankseg: a Consistent Ranking-Based Framework for Segmentation  
*♦Ben Dai<sup>1</sup> and Chunlin Li<sup>2</sup>. <sup>1</sup>The Chinese University of Hong Kong <sup>2</sup>The University of Minnesota*
- 10:50 Multilayer Random Dot Product Graphs: Nonparametric Estimation and Online Change Point Detection  
*Fan Wang<sup>1</sup>, Wanshan Li<sup>2</sup>, Oscar Madrid<sup>3</sup>, ♦Yi Yu<sup>1</sup> and Alessandro Rinaldo<sup>2</sup>. <sup>1</sup>University of Warwick <sup>2</sup>Carnegie Mellon University <sup>3</sup>UCLA*
- 11:15 Floor Discussion.

### Session 23CHI28: Recent Developments on Machine Learning and Precision Medicine

Room: CHONGMING HALL

Organizer: Yufeng Liu, University of North Carolina.  
 Chair: Xiaoling Lu, Renmin University of China.

- 10:00 Multivariate Change Point Detection for Heterogeneous Series  
*Yuxuan Guo<sup>1</sup>, Ming Gao<sup>2</sup> and ♦Xiaoling Lu<sup>1</sup>. <sup>1</sup>Renmin University of China <sup>2</sup>The University of Chicago*
- 10:25 Simultaneous Change Point Detection and Identification for High Dimensional Linear Models  
*♦Bin Liu<sup>1</sup>, Xinsheng Zhang<sup>1</sup> and Yufeng Liu<sup>2</sup>. <sup>1</sup>Department of Statistics and Data Science, School of Management, Fudan University <sup>2</sup>Department of Statistics and Operations Research, Department of Genetics and Department of Biostatistics, University of North Carolina at Chapel Hill*
- 10:50 Locally Weighted Nearest Neighbor Classifier  
*♦Guan Yu<sup>1</sup> and Xingye Qiao<sup>2</sup>. <sup>1</sup>University of Pittsburgh <sup>2</sup>Binghamton University*
- 11:15 Efficient Learning of Optimal Individualized Treatment Rules  
*♦Weibin Mo<sup>1</sup> and Yufeng Liu<sup>2</sup>. <sup>1</sup>Purdue University <sup>2</sup>University of North Carolina at Chapel Hill*
- 11:40 Floor Discussion.

### Session 23CHI30: Recent Developments in Biostatistics with their Applications in Cancer Genomics, Screening and Graph Modeling

Room: NANSHI HALL

Organizer: Judy Zhong, Professor, Division of Biostatistics, Department of Population Health, NYU Grossman School of Medicine.  
 Chair: Judy Zhong, Professor, Division of Biostatistics, Department of Population Health, NYU Grossman School of Medicine.

- 10:00 Bias Correction Models for Ehr Data in the Presence of Non-Random Sampling  
*Judy Zhong.*
- 10:25 Assessing Screening Efficacy in the Presence of Cancer Overdiagnosis  
*♦Ying Huang and Ziding Feng. Fred Hutchinson Cancer Center*
- 10:50 Learning Directed Acyclic Graphs for Ligands and Receptors Based on Spatially Resolved Transcriptomic Analysis  
*Pei Wang. Icahn School of Medicine at Mount Sinai, New York, NY*
- 11:15 Learning Directed Acyclic Graphs with Mixed Variables  
*♦Jie Peng<sup>1</sup>, Pei Wang<sup>2</sup> and Shrabanti Chowdhury<sup>2</sup>. <sup>1</sup>University of California, Davis <sup>2</sup>Icahn School of Medicine at Mount Sinai*
- 11:40 Floor Discussion.

### Session 23CHI31: New Developments in Missing Data Problems and Related Areas

Room: PUTUO HALL

Organizer: Yukun Liu, East China Normal University.  
 Chair: Yukun Liu, East China Normal University.

- 10:00 Instability of Inverse Probability Weighting Methods and a Remedy for Non-Ignorable Missing Data  
*Pengfei Li<sup>1</sup>, Jing Qin<sup>2</sup> and ♦Yukun Liu<sup>3</sup>. <sup>1</sup>University of Waterloo <sup>2</sup>NIH <sup>3</sup>East China Normal University*
- 10:25 Evaluating Dynamic Conditional Quantile Treatment Effects with Applications in Ridesharing  
*♦Ting Li<sup>1</sup>, Chengchun Shi<sup>2</sup>, Zhaohua Lu<sup>3</sup>, Yi Li<sup>3</sup> and Hongtu Zhu<sup>4</sup>. <sup>1</sup>Shanghai University of Finance and Economics <sup>2</sup>London School of Economics and Political Science <sup>3</sup>Didi Chuxing <sup>4</sup>University of North Carolina at Chapel Hill*
- 10:50 Multiply Robust Estimation for Two-Part Regression Models with Missing Semicontinuous Response  
*Qiyin Zheng and ♦Chunlin Wang. Xiamen University*
- 11:15 Semiparametric Inference for Quantile Regression in Imbalanced Semi-Supervised Distributed System  
*♦Shuyi Zhang and Yong Zhou. East China Normal University*
- 11:40 Floor Discussion.

### Session 23CHI40: Advanced Statistical Learning Methods for Complex Data

Room: BAOSHAN HALL

Organizer: Guannan Wang, College of William & Mary.  
 Chair: Xiaofei Zhang, Zhongnan University of Economics and Law.

- 10:00 Distillation Decision Tree  
*♦Xuetao Lu and J.jack Lee.*

- 10:25 Clustering Methods for Microbiome Data  
♦*Peng Liu and Zhili Qiao.* Iowa State University
- 10:50 Collaborative Spectral Clustering in Attributed Networks  
*Pengsheng Ji.* Univ. of Georgia
- 11:15 Floor Discussion.

### Session 23CHI41: Recent Development for Causal Inference and Personalized Medicine

Room: HONGKOU HALL

Organizer: Wenbin Lu, North Carolina State University.  
Chair: Chengchun Shi, London School of Economics.

- 10:00 Towards Trustworthy Explanation: On Causal Rationalization  
*Wenbo Zhang<sup>1</sup>, Tong Wu<sup>2</sup>, Yunlong Wang<sup>2</sup>, Yong Cai<sup>2</sup> and Hengrui Cai<sup>1</sup>.* <sup>1</sup>University of California Irvine <sup>2</sup>IQVIA
- 10:25 Multi-Threshold Change Plane Model: Estimation Theory and Applications in Subgroup Identification  
♦*Jialiang Li<sup>1</sup>, Yaguang Li, Baisuo Jin and Michael Kosorok.* <sup>1</sup>National University of Singapore
- 10:50 Dynamic Treatment Regimes with Infinite Horizon and Outcome-Dependent Observation Process  
♦*Xin Chen and Wenbin Lu.*
- 11:15 Floor Discussion.

### Session 23CHI56: Statistical Modeling and Inferences for Complex Data Analysis

Room: CHANGNING HALL

Organizer: Jian Kang, University of Michigan.  
Chair: Jian Kang, University of Michigan.

- 10:00 Statistical Inferences for Complex Dependence of Multi-modal Imaging Data  
♦*Jinyuan Chang<sup>1</sup>, Jing He<sup>1</sup>, Jian Kang<sup>2</sup> and Mingcong Wu<sup>1</sup>.* <sup>1</sup>Southwestern University of Finance and Economics <sup>2</sup>University of Michigan
- 10:25 A Simultaneous Likelihood Test for Joint Mediation Effects of Multiple Mediators  
♦*Wei Hao and Peter Song.* University of Michigan
- 10:50 Bi-Directional Clustering via Averaged Mixture of Finite Mixtures  
♦*Guanyu Hu<sup>1</sup>, Tianyu Pan<sup>2</sup> and Weinig Shen<sup>1</sup>.* <sup>1</sup>University of Missouri Columbia <sup>2</sup>University of California Irvine
- 11:15 A Simple and Robust Model for Enrollment Projection in Clinical Trials  
*Xiaoxi Zhang and Bo Huang.* Pfizer Inc.
- 11:40 Floor Discussion.

### Session 23CHI66: Advances in Theory and Statistical Applications of Random Fields

Room: XUHUI HALL

Organizer: Juan Du, Kansas State University.  
Chair: Juan Du, Kansas State University.

- 10:00 Almost-Sure Path Properties of Operator Fractional Brownian Motion  
*Wensheng Wang.* Hangzhou Dianzi University
- 10:25 Multivariate Spatial Modeling Based on Conditionally Negative Definite Functions  
♦*Juan Du<sup>1</sup> and Xiaoxi Li<sup>2</sup>.* <sup>1</sup>Kansas State University <sup>2</sup>Kansas State univeristy
- 10:50 Error Bounds for the Measurement of Random Fields and the Relation to the Statistical Model  
*Tianshi Lu.* Wichita State University, Kansas, USA
- 11:15 Statistical Analysis of Multivariate Gaussian Random Fields  
*Yimin Xiao.*
- 11:40 Floor Discussion.

### Session 23CHI67: Recent Statistical Methodological Developments in Genomics

Room: FENGXIAN HALL

Organizer: Yi Xiong, University of Manitoba.  
Chair: Yi Xiong, University of Manitoba.

- 10:00 Zero-inflated Poisson Models with Measurement Error in the Response  
♦*Qihuang Zhang<sup>1</sup> and Grace Y. Yi<sup>2</sup>.* <sup>1</sup>McGill University <sup>2</sup>University of Western Ontario
- 10:25 Neural Network Models for Sequence-Based Tcr and Hla Association Prediction  
♦*Si Liu, Philip Bradley and Wei Sun.* Fred Hutchinson Cancer Center
- 10:50 Association Analysis Between the t-Cell Receptor Repertoire and Clinical Phenotypes  
*Meiling Liu<sup>1</sup>, Juna Goo<sup>2</sup>, Yang Liu<sup>3</sup>, Wei Sun<sup>1</sup>, Michael Wu<sup>1</sup>, Li Hsu<sup>1</sup> and Qianchuan He<sup>1</sup>.* <sup>1</sup>Fred Hutchinson Cancer Center <sup>2</sup>Boise State University <sup>3</sup>Wright State University
- 11:15 Knockofftrio: a Knockoff Framework for the Identification of Putative Causal Variants in Genome-Wide Association Studies with Trio Design  
♦*Yi Yang<sup>1</sup>, Chen Wang<sup>2</sup>, Linxi Liu<sup>3</sup>, Joseph Buxbaum<sup>4</sup>, Zihua He<sup>5</sup> and Iuliana Ionita-Laza<sup>2</sup>.* <sup>1</sup>City University of Hong Kong <sup>2</sup>Columbia University <sup>3</sup>University of Pittsburgh <sup>4</sup>Icahn School of Medicine at Mount Sinai <sup>5</sup>Stanford University
- 11:40 Floor Discussion.

### Session 23CHI8: Study Design and Statistical Considerations in Oncology Development: Methodology and Case Sharing

Room: JIADING HALL

Organizer: Qian Jiang, Department of Biostatistics and Programming, China, Sanofi, Inc..  
Chair: Qian JIANG, Department of Biostatistics and Programming, China, Sanofi, Inc..

- 10:00 Design Considerations for Early-Phase Clinical Trials of Car-t Therapies  
*Wentian Guo.* Astrazeneca R&D China
- 10:25 Statistical and Operational Consideration in Umbrella Trial -a Real Case Study  
♦*Monica Li, Keisuke Tada and Rick Zhang.*

- 10:50 An Oncology Case Example of Dose Optimization using the Pick-the-Winner Approach  
 ♦*Liang Zhao*<sup>1</sup>, *Qiuyan Wang*<sup>1</sup> and *Gu Mi*<sup>2</sup>. <sup>1</sup>Department of Biostatistics and Programming, China, Sanofi, Inc.  
<sup>2</sup>Department of Biostatistics and Programming, US, Sanofi, Inc.
- 11:15 Adjusting Survival Time Estimates Accounting for Treatment Switching  
 ♦*Yujuan Gao, Songzi Li and Jiang Li*. BeiGene, Ltd.
- 11:40 Floor Discussion.

### Session 23CHI97: Dealing with Missing Data: Recent Methodological Advances

Room: JINSHAN HALL  
 Organizer: Zhengyuan Zhu, Iowa State University.  
 Chair: Emily Berg, Iowa State University.

- 10:00 Doubly Robust Estimators for Generalizing Treatment Effects on Survival Outcomes from Randomized Controlled Trials to a Target Population  
*Dasom Lee*<sup>1</sup>, ♦*Shu Yang*<sup>1</sup> and *Xiaofei Wang*<sup>2</sup>. <sup>1</sup>North Carolina State University <sup>2</sup>Duke University
- 10:25 Doubly Robust Estimation with Outliers under Missing at Random  
 ♦*Jae-Kwang Kim*<sup>1</sup>, *Jeongsup Han*<sup>2</sup>, *Hengfang Wang*<sup>3</sup> and *Youngjo Lee*<sup>2</sup>. <sup>1</sup>Iowa State University <sup>2</sup>Seoul National University <sup>3</sup>Fujian Normal University
- 10:50 Marginal Treatment Effect Estimation without Ignorability using Observational Study  
*Guoliang Ma*<sup>1</sup> and ♦*Cindy Yu*<sup>2</sup>. <sup>1</sup>Iowa State University <sup>2</sup>Iowa State Univestiy
- 11:15 Floor Discussion.

## July 2, 13:00-14:40

### Session 23CHI103: Modern Topics in Design of Experiments (DOE)

Room: JINGAN HALL  
 Organizer: Wei Zheng, University of Tennessee.  
 Chair: Wei Zheng, wzheng9@utk.edu.

- 13:00 Thompson Sampling with Discrete Prior  
*Xueru Zhang*<sup>1</sup>, *Lan Gao*<sup>2</sup> and ♦*Wei Zheng*<sup>2</sup>. <sup>1</sup>Purdue University <sup>2</sup>University of Tennessee
- 13:25 A Maximin p-Efficient Design for Multivariate Generalized Linear Models  
 ♦*Yiou Li*<sup>1</sup>, *Lulu Kang*<sup>2</sup> and *Xinwei Deng*<sup>3</sup>. <sup>1</sup>Associate Prof. DePaul University <sup>2</sup>Associate Prof. Illinois Institute of Technology <sup>3</sup>Prof. Virginia Tech
- 13:50 A New and Flexible Design Construction for Orthogonal Arrays for Modern Applications  
*Yuanzhen He*<sup>1</sup>, ♦*C. Devon Lin*<sup>2</sup> and *Fasheng Sun*<sup>3</sup>. <sup>1</sup>School of Statistics, Beijing Normal University <sup>2</sup>Department of Mathematics and Statistics, Queen's University <sup>3</sup>KLAS and School of Mathematics and Statistics, Northeast Normal University

- 14:15 Maximum One-Factor-at-a-Time Designs for Screening in Computer Experiments  
 ♦*Qian Xiao*<sup>1</sup>, *Roshan Joseph*<sup>2</sup> and *Douglas Ray*<sup>3</sup>. <sup>1</sup>University of Georgia <sup>2</sup>Georgia Institute of Technology <sup>3</sup>US Army DEVCOM Armaments Center
- 14:40 Floor Discussion.

### Session 23CHI104: Applications of Modern Statistical Methods for High-Dimensional and Complex Data

Room: NANHUI HALL  
 Organizer: Wenbo Wu, The University of Texas at San Antonio.  
 Chair: Wenbo Wu, The University of Texas at San Antonio.

- 13:00 Bayesian Borrowing with Multiple Heterogeneous Historical Studies using Order Restricted Normalized Power Prior  
*Zifei Han*. University of International Business and Economics
- 13:25 On Sufficient Variable Screening using log Odds Ratio Filter  
*Baoying Yang*, ♦*Wenbo Wu* and *Xiangrong Yin*.
- 13:50 Bayesian Estimation of Malware Detection Metrics without Knowing Ground Truth  
*Ambassador Negash*<sup>1</sup>, *Zifei Han*<sup>2</sup>, *Min Wang*<sup>3</sup>, *Shouhuai Xu*<sup>4</sup> and ♦*Keying Ye*<sup>3</sup>. <sup>1</sup>Footlock Inc. <sup>2</sup>University of International Business and Economics <sup>3</sup>University of Texas at San Antonio <sup>4</sup>University of Colorado at Colorado Springs
- 14:15 Exploring the Causal Relationship Between Geriatric Depression and Alzheimer's Disease  
 ♦*Yuxia Zhang*<sup>1</sup>, *Yubai Yuan*<sup>2</sup>, *Fei Xue*<sup>3</sup>, *Qi Xu*<sup>4</sup> and *Annie Qu*<sup>4</sup>. <sup>1</sup>The University of Texas at San Antonio <sup>2</sup>The Pennsylvania State University <sup>3</sup>Purdue University <sup>4</sup>University of California, Irvine
- 14:40 Floor Discussion.

### Session 23CHI105: Survival Analysis and Quantile Regression

Room: SONGJIANG HALL  
 Organizer: Xuerong Chen, Southwestern University of Finance and Economics.  
 Chair: Xuerong Chen, Southwestern University of Finance and Economics.

- 13:00 From Conditional Quantile Regression to Marginal Quantile Estimation with Applications to Missing Data and Causal Inference  
 ♦*Huijuan Ma, Jing Qin and Yong Zhou*.
- 13:25 Default Risk Prediction and Feature Extraction using a Penalized Deep Neural Network  
 ♦*Cunjie Lin*<sup>1</sup>, *Nan Qiao*<sup>1</sup>, *Wenli Zhang*<sup>1</sup>, *Yang Li*<sup>1</sup> and *Shuangge Ma*<sup>2</sup>. <sup>1</sup>School of Statistics, Renmin University of China <sup>2</sup>Department of Biostatistics, Yale University
- 13:50 Neural Network on Interval Censored Data with Application to the Prediction of Alzheimer's Disease  
 ♦*Tao Sun*<sup>1</sup> and *Ying Ding*<sup>2</sup>. <sup>1</sup>Renmin University of China <sup>2</sup>University of Pittsburgh
- 14:15 Floor Discussion.

## Session 23CHI106: Statistical Inference with Large Scale Data

Room: JINSHAN HALL

Organizer: Qihua Wang, Chinese Academy of Sciences, Academy of Mathematics and Systems Science.

Chair: Qihua Wang, Qihua Wang.

13:00 Integrative Conformal p-Values for Out-of-Distribution Testing with Labeled Outliers

*Wenguang Sun.* Zhejiang University

13:25 去中心化数据分析中的若干统计学问题

*Weidong Liu.* Shanghai Jiao Tong University

13:50 A Robust Fusion-Extraction Procedure with Summary Statistics in the Presence of Biased Sources

*Ruoyu Wang<sup>1</sup>, ♦Qihua Wang<sup>1</sup> and Wang Miao<sup>2</sup>.* <sup>1</sup>Academy of Mathematics and Systems Science, Chinese Academy of Sciences <sup>2</sup>Peking University

14:15 Sequential Data Integration under Dataset Shift

♦*Ying Sheng<sup>1</sup>, Jing Qin<sup>2</sup> and Chiung-Yu Huang<sup>3</sup>.*

<sup>1</sup>Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences <sup>2</sup>National Institute of Allergy and Infectious Diseases, National Institute of Health <sup>3</sup>Department of Epidemiology & Biostatistics, University of California at San Francisco

14:40 Floor Discussion.

## Session 23CHI108: Joint Modeling of Multiple Types of Data in Health Studies

Room: BAOSHAN HALL

Organizer: Xinyue Li, School of Data Science, City University of Hong Kong, Hong Kong SAR, China.

Chair: Xinyue Li, School of Data Science, City University of Hong Kong, Hong Kong SAR, China.

13:00 Joint Modeling of Survival, Longitudinal and Recurrent Event Data for Individual Electronic Health Records: An Application in Community-Dwelling Elderly Surveillance

♦*Hao Pan<sup>1</sup>, Xinyue Li<sup>2</sup>, Yang Zhao<sup>3</sup>, Hailiang Wang<sup>4</sup> and Kwok Leung Tsui<sup>5</sup>.* <sup>1</sup>Shanghai Jiao Tong University School of Medicine, Shanghai Children's Medical Center <sup>2</sup>City University of Hong Kong, School of Data Science <sup>3</sup>Sun Yat-sen University School of Public Health (Shenzhen) <sup>4</sup>The Hong Kong Polytechnic University, School of Design <sup>5</sup>Virginia Tech, Department of Industrial and Systems Engineering

13:25 Pathological Imaging-Assisted Cancer Gene-Environment Interaction Analysis

*Kuangnan Fang<sup>1</sup>, Jingmao Li<sup>1</sup>, Qingzhao Zhang<sup>2</sup>, ♦Yaqing Xu<sup>3</sup> and Shuangge Ma<sup>4</sup>.* <sup>1</sup>Department of Statistics and Data Science, School of Economics, Xiamen University, Xiamen <sup>2</sup>The Wang Yanan Institute for Studies in Economics, Xiamen University <sup>3</sup>School of Public Health, Shanghai Jiao Tong University School of Medicine <sup>4</sup>Department of Biostatistics, Yale School of Public Health

13:50 Human Disease Clinical Treatment Network for the Elderly: Analysis of the Medicare Inpatient Length of Stay and Readmission Data

♦*Hao Mei<sup>1</sup>, Ruofan Jia<sup>2</sup>, Guanzhong Qiao<sup>3</sup>, Zhenqiu Lin<sup>4</sup> and Shuangge Ma<sup>5</sup>.* <sup>1</sup>School of Statistics, Renmin University of China <sup>2</sup>The Wang Yanan Institute for Studies in Economics, Xiamen University <sup>3</sup>Department of Orthopaedic,

The First Hospital of Tsinghua University <sup>4</sup>Center for Outcomes Research and Evaluation, Yale-New Haven Hospital

<sup>5</sup>Department of Biostatistics, Yale University

14:15 Functional Adaptive Double-Sparsity Estimator for Functional Linear Regression Model with Multiple Functional Covariates

♦*Cheng Cao<sup>1</sup>, Jiguo Cao<sup>2</sup>, Hailiang Wang<sup>3</sup>, Kwok-Leung Tsui<sup>4</sup> and Xinyue Li<sup>1</sup>.* <sup>1</sup>City University of Hong Kong

<sup>2</sup>Simon Fraser University <sup>3</sup>The Hong Kong Polytechnic University <sup>4</sup>Virginia Polytechnic Institute and State University

14:40 Floor Discussion.

## Session 23CHI113: Advanced Analytic and Innovation in Healthcare

Room: LONGFENG HALL-II

Organizer: Sammi Tang, Servier pharmaceutical.

Chair: Sammi Tang, Servier pharmaceuticals.

13:00 Recent Development for Ai/MI for Drug Discovery

*Haoda Fu.* Eli Lilly and Company

13:25 A Digital Health Tool for Dynamic Patient Recruitment Prediction in Multicenter Clinical Trials

*Jianmin Chen<sup>1</sup>, Zhaowei Hua<sup>2</sup>, Qian Meng<sup>2</sup>, Truffaut-Chalet Luice, ♦Zhaoyang Teng<sup>2</sup> and Sammi Tang<sup>2</sup>.* <sup>1</sup>University of Connecticut <sup>2</sup>Servier Pharmaceuticals

13:50 Data Simulation to Forecast the Outcomes of the Favor Iii China Trial

*Yang Wang.* National Center for Cardiovascular Diseases

14:15 To be Provided

*Rui (Sammi) Tang.*

14:40 Bayesian Learning of Covid-19 Vaccine Safety While Incorporating Adverse Events Ontology

*Bangyao Zhao, Yuan Zhong, ♦Jian Kang and Lili Zhao.* University of Michigan

Floor Discussion.

## Session 23CHI116: Development on Factorial Designs

Room: PUTUO HALL

Organizer: Shengli Zhao, Qufu Normal University.

Chair: Shengli Zhao, Qufu Normal University.

13:00 r-Optimal Designs for q-Ingredient Becker's Models for Experiments with Mixtures

♦*Chongqi Zhang<sup>1</sup> and Junpeng Li<sup>2</sup>.* <sup>1</sup>Professor <sup>2</sup>Mr

13:25 The Construction and Properties of a Class of Clear Compromise Designs

♦*Qi Zhou and Xue Yang.* School of Statistics, Tianjin University of Finance and Economics

13:50 Research on Split-Plot Designs under the General Minimum Lower-Order Confounding Criterion

♦*Tao Sun and Shengli Zhao.* Qufu Normal University

14:15 The Development of General Minimum Lower-Order Confounding Criterion in the Design of Experiments

*Zhiming Li.* Xinjiang University

14:40 Floor Discussion.

**Session 23CHI117: Stochastic Modeling of Complex Data**

Room: FENGXIAN HALL

Organizer: Shuyang Bai, University of Georgia.

Chair: Ting Zhang, University of Georgia.

13:00 Joint Sum-Max Limit for a Class of Long-Range Dependent Processes with Heavy Tails

*Shuyang Bai and ♦He Tang.* University of Georgia

13:25 How to Identify your Most Valuable Customers

*Chen Xing.* Zeta Technologies

13:50 Influence Effects in Social Networks: Inward and Outward Spillovers of One Unit's Treatment

*♦Fei Fang and Laura Forastiere.* Yale University

14:15 High-Quantile Regression for Tail-Dependent Time Series

*Ting Zhang.* University of Georgia

14:40 Floor Discussion.

**Session 23CHI119: Advanced Statistical Methods for Data Analysis**

Room: JIADING HALL

Organizer: Linlin Dai, Southwestern University Of Finance And Economics.

Chair: Feifei Guo, Beijing Institute of Technology.

13:00 Combining Extreme Value Theory with Martingale Regression in Market Risk Analytics and Portfolio Management.

*♦Wei Dai and Tze Leung Lai.*

13:25 A Clt for the Lss of Large Dimensional Sample Covariance Matrices with Diverging Spikes

*♦Zhijun Liu, Jiang Hu, Zhidong Bai and Haiyan Song.*  
Northeast Normal University

13:50 Inference in Nonstationary Heavy-Tailed Ar Process via Model Selection

*♦Feifei Guo<sup>1</sup> and Rui She<sup>2</sup>.* <sup>1</sup>Beijing Institute of Technology <sup>2</sup>Southwestern University of Finance and Economics

14:15 A Weighted Average Distributed Estimator for High Dimensional Parameter

*♦Jun Lu<sup>1</sup>, Mengyao Li<sup>2</sup> and Chenping Hou<sup>1</sup>.* <sup>1</sup>National University of Defense Technology <sup>2</sup>Xi'an Jiaotong University

14:40 Discussant: Feifei Guo, Beijing Institute of Technology

Floor Discussion.

**Session 23CHI122: Addressing Computational Challenges in Analyzing High-Throughput Genomic and Epigenetics Data**

Room: LUWAN HALL

Organizer: Ziyi Li, The University of Texas MD Anderson Cancer Center.

Chair: Zhenxing Guo, Chinese University of Hong Kong, Shenzhen (CUHK-SZ).

13:00 G3dc: a Gene-Graph-Guided Selective Deep Clustering Method for Single Cell Rna-Seq Data

*Shuqing He, Jicong Fan and ♦Tianwei Yu.*

13:25 Supervised Cell Type Identification for Single Cell Atac-Seq Data

*Hao Wu.* Shenzhen Institute of Advanced Technology

13:50 Evaluation of Epitranscriptome-Wide N6-Methyladenosine Differential Analysis Methods

*Daoyu Duan<sup>1</sup>, Wen Tang<sup>1</sup>, Runshu Wang<sup>2</sup>, ♦Zhenxing Guo<sup>3</sup> and Hao Feng<sup>1</sup>.* <sup>1</sup>Department of Population and Quantitative Health Sciences, Case Western Reserve University <sup>2</sup>Department of Biostatistics, University of Michigan<sup>3</sup>School of Data Science, The Chinese University of Hong Kong - Shenzhen

14:15 Rationally Design Generative-Adversarial Models for Delimiting the Regulatory Map in Silico

*Ge Gao.* Peking University

14:40 Floor Discussion.

**Session 23CHI171: Advances in Statistical Methods for Biomedical Studies**

Room: T-1

Organizer: Zhezhen Jin, Columbia University.

Chair: Zhezhen Jin, Columbia University.

13:00 On the Instrumental Variable Estimation with many Weak and Invalid Instruments

*♦Yiqi Lin<sup>1</sup>, Frank Windmeijer<sup>2</sup>, Xinyuan Song<sup>1</sup> and Qingliang Fan<sup>3</sup>.* <sup>1</sup>Dept of Stat, The Chinese University of Hong Kong <sup>2</sup>Dept of Stat, University of Oxford <sup>3</sup>Dept of Econ, The Chinese University of Hong Kong

13:25 A Flexible Bayesian Clustering of Dynamic Subpopulations in Neural Spiking Activity

*Ganchao Wei, Ian Stevenson and ♦Xiaojing Wang.* University of Connecticut

13:50 A Knockoff Framework for Effective Biomarker Identification in Drug Development: With Application to a Psoriatic Arthritis Clinical Trial

*Matthias Kormáksson, Kostas Sechidis, Xuan Zhu, David Ohlssen and ♦Cong Zhang.* Novartis

14:15 On the Instrumental Variable Estimation with many Weak and Invalid Instruments

*Yiqi LIN.*

14:40 Floor Discussion.

**Session 23CHI32: Advances Developments in Statistical Methodology**

Room: NANSHI HALL

Organizer: Baoying Yang, Southwest Jiaotong University.

Chair: Baoying Yang, Southwest Jiaotong University.

13:00 Efficient Estimation of Cox Model with Random Change Point

*Yalu Ping<sup>1</sup>, ♦Xuerong Chen<sup>1</sup> and Jianguo Sun<sup>2</sup>.* <sup>1</sup>Southwestern University of Finance and Economics <sup>2</sup>University of Missouri

13:25 Nearest-Neighbor Sampling Based Conditional Independence Testing

*Shuai Li<sup>1</sup>, ♦Ziqi Chen<sup>1</sup>, Hongtu Zhu<sup>2</sup>, Dan Wang<sup>3</sup> and Wang Wen<sup>4</sup>.* <sup>1</sup>East China Normal University <sup>2</sup>The University of North Carolina at Chapel Hill <sup>3</sup>New York University Shanghai <sup>4</sup>Central South University

- 13:50 Mean Change Point Detection Based on Jump Information Criterion  
*Zhiming Xia.* Northwest University
- 14:15 Floor Discussion.

### Session 23CHI33: Recent Statistical Developments for Complex High-Dimensional Data

Room: XUHUI HALL  
Organizer: Lu Tang, University of Pittsburgh.  
Chair: Ling Zhou, Southwestern University of Finance and Economics.

- 13:00 Identification of Prognostic and Predictive Subgroups for Clustered Survival Data  
*Ye He<sup>1</sup>, Dongsheng Tu<sup>2</sup> and ♦Ling Zhou<sup>3</sup>.* <sup>1</sup>Sichuan Normal University <sup>2</sup>Queen's University <sup>3</sup>Southwestern University of Finance and Economics
- 13:25 A Variational Bayesian Approach to Identifying Whole-Brain Directed Networks with Fmri Data  
*Yaotian Wang<sup>1</sup>, Guofen Yan<sup>2</sup>, Xiaofeng Wang<sup>3</sup>, Shuoran Li<sup>1</sup>, Lingyi Peng<sup>1</sup>, Dana Tudorascu<sup>1</sup> and ♦Tingting Zhang<sup>1</sup>.* <sup>1</sup>University of Pittsburgh <sup>2</sup>University of Virginia <sup>3</sup>Cleveland Clinic
- 13:50 d-Gcca: Decomposition-Based Generalized Canonical Correlation Analysis for Multi-View High-Dimensional Data  
♦*Hai Shu<sup>1</sup>, Zhe Qu<sup>2</sup> and Hongtu Zhu<sup>3</sup>.* <sup>1</sup>New York University <sup>2</sup>Tulane University <sup>3</sup>The University of North Carolina at Chapel Hill
- 14:15 Floor Discussion.

### Session 23CHI36: Statistical Machine Learning and Complex Life Time Data Analysis

Room: LONGFENG HALL-I  
Organizer: Kaida Cai, Southeast University.  
Chair: Kaida Cai, Southeast University.

- 13:00 Identification of Survival Relevant Genes with Measurement Error in Gene Expression Incorporated  
♦*Juan Xiong<sup>1</sup> and Wenqing He<sup>2</sup>.* <sup>1</sup>Shenzhen University <sup>2</sup>Western University
- 13:25 A Novel Approach to Ordinal Classification with Deep Neural Networks  
♦*Yiwei Fan<sup>1</sup>, Xiaoshi Lu<sup>2</sup> and Xiaoling Lu<sup>2</sup>.* <sup>1</sup>School of Mathematics and Statistics, Beijing Institute of Technology <sup>2</sup>School of Statistics, Renmin University of China
- 13:50 Supervised Topic Modeling: Optimal Estimation and Statistical Inference  
♦*Ruijia Wu<sup>1</sup>, Linjun Zhang<sup>2</sup> and T. Tony Cai<sup>3</sup>.* <sup>1</sup>Shanghai Jiao Tong University <sup>2</sup>Rutgers University <sup>3</sup>University of Pennsylvania
- 14:15 Multistate Modeling and Structure Selection for Multitype Recurrent Events and Terminal Event Data  
*Chuoxin Ma.* Beijing Normal University-Hong Kong Baptist University United International College (UIC)
- 14:40 Floor Discussion.

### Session 23CHI37: Statistical Challenges for Analyzing Complex Data

Room: LONGFENG HALL-III  
Organizer: Li-Shan Huang, National Tsing Hua University.  
Chair: Min Zhou, Beijing Normal University-Hong Kong Baptist University United International College, CHINA.

- 13:00 Semiparametric Efficient Estimation of Genetic Relatedness with Double Machine Learning  
*Xu Guo<sup>1</sup>, Yiyuan Qian<sup>1</sup>, Hongwei Shi<sup>1</sup>, Weichao Yang<sup>1</sup> and ♦Niwen Zhou<sup>2</sup>.* <sup>1</sup>School of Statistics, Beijing Normal University <sup>2</sup>Advanced Institute of Natural Sciences, Beijing Normal University
- 13:25 Causal Inference Methods for Multiple Treatment Group Evaluations  
*Hongwei Zhao.* Texas A&M University
- 13:50 Unifying Estimation and Inference for Linear Regression with Stationary and Integrated or Near-Integrated Variables  
*Shaoxin Hong<sup>1</sup>, Daniel Henderson<sup>2</sup>, ♦Jiancheng Jiang<sup>3</sup> and Qingshan Ni<sup>4</sup>.* <sup>1</sup>Shandong University <sup>2</sup>University of Alabama <sup>3</sup>University of North Carolina at Charlotte <sup>4</sup>Hunan University
- 14:15 Bolt-Ssi: a Statistical Approach to Screening Interaction Effects for Ultra-High Dimensional Data  
♦*Min Zhou<sup>1</sup>, Mingwei Dai<sup>2</sup>, Yuan Yao<sup>3</sup>, Jin Liu<sup>4</sup>, Can Yang<sup>5</sup> and Heng Peng<sup>6</sup>.* <sup>1</sup>Beijing Normal University-Hong Kong Baptist University United International College <sup>2</sup>Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics <sup>3</sup>Victoria University of Wellington, School of Mathematics and Statistics <sup>4</sup>Duke-NUSS Graduate Medical School <sup>5</sup>The Hong Kong University of Science and Technology <sup>6</sup>Hong Kong Baptist University
- 14:40 Floor Discussion.

### Session 23CHI39: Recent Advances on Covariate Models and Applications

Room: HONGKOU HALL  
Organizer: Jingjing Wu, University of Calgary.  
Chair: Jingjing Wu, University of Calgary.

- 13:00 Low-Tubal-Rank Tensor Sensing and Robust Pca from Quantized Measurements  
*Jianjun Wang.*
- 13:25 Reseach on Consumer Purchase Intention Based on Short Video Platform  
♦*Bo Li<sup>1</sup>, Xinghui Xiao<sup>2</sup> and Chen Wang<sup>1</sup>.* <sup>1</sup>Communication University of China <sup>2</sup>Beijing Vocational College of Electronic Science and Technology
- 13:50 Limit Behaviours for Nonlinear Regression Models  
*Yu Miao.* Henan Normal University
- 14:15 Minimum Profile Hellinger Distance Estimation of Covariate Models  
♦*Bowei Ding<sup>1</sup>, Rohana J. Karunamuni<sup>2</sup> and Jingjing Wu<sup>1</sup>.* <sup>1</sup>Department of Mathematics and Statistics, University of Calgary <sup>2</sup>Department of Mathematical and Statistical Sciences, University of Alberta
- 14:40 Floor Discussion.

**Session 23CHI46: Advances in Network and Complex Data Analysis**

Room: CHANGNING HALL

Organizer: Pengsheng Ji, Univ. of Georgia.

Chair: Pengsheng Ji, Univ. of Georgia.

13:00 Sparse Gaussian Network Model for Single-Cell Rna-Seq Data

*Mingjin Liu<sup>1</sup>, Susmita Datta<sup>1</sup> and ♦Yang Yang<sup>2</sup>.*  
<sup>1</sup>University of Florida <sup>2</sup>University of Georgia

13:25 Nonparametric Regression for 3d Point Cloud Learning

*♦Xinyi Li<sup>1</sup>, Shan Yu, Yueying Wang, Guannan Wang, Li Wang and Ming-Jun Lai.* <sup>1</sup>Clemson University

13:50 Nonparametric Link Prediction for Networks and Bipartite Graph

*Jia Shen Lu and ♦Kehui Chen.* University of Pittsburgh, USA

14:15 Floor Discussion.

**July 2, 15:00-16:40****Session 23CHI109: Survival Analysis**

Room: LONGFENG HALL-I

Organizer: Jialiang Li, National University of Singapore, Singapore.

Chair: Jialiang Li, National University of Singapore, Singapore.

15:00 A Weighted Generalized Win-Odds Regression Model for Composite Endpoints

*Bang Wang and ♦Yu Cheng.* University of Pittsburgh

15:25 Kernel Meets Sieve: Transformed Hazards Models with Sparse Longitudinal Covariates

*Dayu Sun<sup>1</sup>, Zhuowei Sun<sup>2</sup>, Xinqiu Zhao<sup>3</sup> and ♦Hongyuan Cao<sup>4</sup>.* <sup>1</sup>Emory University <sup>2</sup>Jilin University <sup>3</sup>Hong Kong Polytechnic University <sup>4</sup>Florida State University

15:50 Assessing Dynamic Covariate Effects with Survival Data

*Ying Cui and ♦Limin Peng.* Emory University

16:15 Doubly Robust Estimation under Covariate-Induced Dependent Left Truncation

*Yuyao Wang<sup>1</sup>, Andrew Ying<sup>2</sup> and ♦Ronghui Xu<sup>1</sup>.*  
<sup>1</sup>University of California, San Diego <sup>2</sup>Google Inc.

16:40 Floor Discussion.

**Session 23CHI120: Advance in Statistical Methods for Large and Complex Data**

Room: CHANGNING HALL

Organizer: Dehan Kong, University of Toronto.

Chair: Dehan Kong, University of Toronto.

15:00 Nonlinear Expectation for Bandit Learning

*Xiaodong Yan.* Shandong University

15:25 Functional Linear Operator Quantile Regression for Sparse Longitudinal Data

*♦Xingcai Zhou<sup>1</sup>, Tingyu Lai<sup>2</sup> and Linglong Kong<sup>3</sup>.*  
<sup>1</sup>Nanjing Audit University <sup>2</sup>Guangxi Normal University  
<sup>3</sup>University of Alberta

15:50 Order Statistics Approaches to Unobserved Heterogeneity in Auctions

*Yao Luo<sup>1</sup>, ♦Peijun Sang<sup>2</sup> and Ruli Xiao<sup>3</sup>.* <sup>1</sup>University of Toronto <sup>2</sup>University of Waterloo <sup>3</sup>Indiana University

16:15 Significance Test for Multinomial Naive Bayes Classifier with Ultra-High Dimensional Binary Features

*♦Baiguo An, Juan Zhang, Beibei Zhang and Wenliang Pan.*

16:40 Floor Discussion.

**Session 23CHI121: Modern Statistical and Machine Learning Methods for Analysis of -Omics Data**

Room: JINGAN HALL

Organizer: Qi Long, University of Pennsylvania.

Chair: Qi Long, University of Pennsylvania.

15:00 Integrative Analysis of Gene Expression and Histology Images in Spatial Transcriptomics

*♦Mingyao Li and Daiwei Zhang.* University of Pennsylvania

15:25 Bayesian Pathway Analysis over Brain Network Mediators and Genetic Exposure for Survival Data

*Xinyuan Tian<sup>1</sup>, Fan Li<sup>1</sup>, Li Shen<sup>2</sup>, Denise Esserman<sup>1</sup> and ♦Yize Zhao<sup>1</sup>.* <sup>1</sup>Yale University <sup>2</sup>University of Pennsylvania

15:50 Synthetic Rna Sequencing Data from Pilot Studies using Deep Generative Models

*♦Yunhui Qi<sup>1</sup>, Xinyi Wang<sup>2</sup> and Lixuan Qin<sup>3</sup>.* <sup>1</sup>Iowa State University <sup>2</sup>Columbia University <sup>3</sup>Memorial Sloan Kettering Cancer Center

16:15 Differential Inference for Single-Cell Rna-Sequencing Data

*♦Fangda Song<sup>1</sup>, Kelvin Y. Yip<sup>2</sup> and Yingying Wei<sup>2</sup>.* <sup>1</sup>The Chinese University of Hong Kong, Shenzhen <sup>2</sup>The Chinese University of Hong Kong

16:40 Floor Discussion.

**Session 23CHI125: Large Dimensional Random Matrix and Its Applications**

Room: NANHUI HALL

Organizer: Shurong Zheng, Northeast Normal University.

Chair: Zhaoyuan Li, The Chinese University of Hong Kong, Shenzhen.

15:00 Block-Diagonal Test for High-Dimensional Covariance Matrices

*Jiayu Lai<sup>1</sup>, ♦Xiaoyi Wang<sup>2</sup>, Shurong Zheng<sup>1</sup> and Kaige Zhao<sup>1</sup>.* <sup>1</sup>Northeast Normal University <sup>2</sup>Beijing Normal University

15:25 Estimating the Number of Communities Based on Individual-Centered Partial Information

*Xiyue Zhu, Xiao Han and ♦Qing Yang.* University of Science and Technology of China

15:50 On the Asymptotic Properties of Spike Eigenvalues and Eigenvectors of Signal-Plus-Noise Matrices with their Applications

*♦Yiming Liu<sup>1</sup>, Zhixiang Zhang<sup>2</sup>, Guangming Pan<sup>3</sup> and Lingyue Zhang.* <sup>1</sup>Jinan University <sup>2</sup>University of Pennsylvania <sup>3</sup>NANYANG TECHNOLOGICAL UNIVERSITY

- 16:15 Random Matrix Methods for Machine Learning: An Application to “lossless”Compression of Deep Neural Networks  
*Zhenyu Liao.* Huazhong University of Science and Technology (HUST)  
 16:40 Floor Discussion.

### Session 23CHI126: Recent Developments in Biostatistics and Beyond

Room: SONGJIANG HALL  
 Organizer: Guanyu Hu, University of Missouri.  
 Chair: Guanyu Hu, University of Missouri.

- 15:00 Staarpipeline: An all-in-One Rare-Variant Tool for Biobank-Scale Whole-Genome Sequencing Data  
 ♦*Zilin Li<sup>1</sup>, Xihao Li<sup>2</sup> and Xihong Lin<sup>2</sup>.* <sup>1</sup>Indiana University School of Medicine <sup>2</sup>Harvard T.H. Chan School of Public Health  
 15:25 Bayesian Sparse Gaussian Mixture Model for Clustering in High Dimensions  
 ♦*Dapeng Yao<sup>1</sup>, Fangzheng Xie<sup>2</sup> and Yanxun Xu<sup>1</sup>.* <sup>1</sup>Johns Hopkins University <sup>2</sup>Indiana University  
 15:50 Prediction-Assisted Candidate Screening with Fdr Control via Conformal Inference  
 ♦*Ying Jin and Emmanuel Candes.* Stanford University  
 16:15 Discussant: Guanyu Hu, University of Missouri Columbia  
 16:40 Floor Discussion.

### Session 23CHI127: Experimental Designs and their Applications (II)

Room: JIADING HALL  
 Organizer: Min-Qian Liu, Nankai University.  
 Chair: Rong-Xian Yue, Shanghai Normal University.

- 15:00 Statistical Inference Through Combining Physical and Computer Experiments  
*Yang Li and ♦Shifeng Xiong.*  
 15:25 A Sequential Design for Determining the Quantile Curve in Two-Factor Sensitivity Tests  
*Yuxia Liu, ♦Yubin Tian and Dianpeng Wang.* Beijing Institute of Technology  
 15:50 A Subsampling Method for Regression Problems Based on Minimum Energy Criterion  
*Wenlin Dai<sup>1</sup>, Yan Song<sup>1</sup> and ♦Dianpeng Wang<sup>2</sup>.* <sup>1</sup>Renmin University of China <sup>2</sup>Beijing institute of technology  
 16:15 Global Likelihood Sampler for Multimodal Distributions  
 ♦*Si-Yu Yi, Ze Liu, Min-Qian Liu and Yong-Dao Zhou.* Nankai University  
 16:40 Floor Discussion.

### Session 23CHI128: Space-Filling Designs (II)

Room: T-1  
 Organizer: Jian-Feng Yang, Nankai University.  
 Chair: Shifeng Xiong, Academy of Mathematics and Systems Science, Chinese Academy of Sciences.

- 15:00 Efficient Kriging using Designs with Low Fill Distance and High Separation Distance  
*Xu He.* Chinese Academy of Sciences  
 15:25 Doubly Coupled Designs for Computer Experiments with both Qualitative and Quantitative Factors  
 ♦*Feng Yang<sup>1</sup>, C. Devon Lin<sup>2</sup>, Yongdao Zhou<sup>3</sup> and Yuanzhen He<sup>4</sup>.* <sup>1</sup>Sichuan Normal University <sup>2</sup>Queen’s University <sup>3</sup>Nankai University <sup>4</sup>Beijing Normal University  
 15:50 Deterministic Construction Methods for Uniform Designs  
 ♦*Liang-Wei Qi, Ze Liu and Yong-Dao Zhou.* School of Statistics and Data Science, Nankai University  
 16:15 Deterministic Construction Methods for Uniform Designs  
 ♦*Liangwei Qi, Ze Liu and Yongdao Zhou.* Nankai University  
 16:40 A Minimum Aberration Type Criterion for Selecting Space-Filling Designs.  
 ♦*Ye Tian<sup>1</sup> and Hongquan Xu<sup>2</sup>.* <sup>1</sup>Beijing University of Posts and Telecommunications <sup>2</sup>University of California, Los Angeles  
 Floor Discussion.

### Session 23CHI133: Statistical and Machine Learning Methods for Biomedical Research

Room: JINSHAN HALL  
 Organizer: Hongyu Zhao, Yale University.  
 Chair: Tao Wang, Shanghai Jiao Tong University.

- 15:00 Fusion of Supervised Learning and Reinforcement Learning for Dynamic Treatment Recommendation  
*Yiyuan Liu<sup>1</sup>, Linjiajie Fang<sup>2</sup>, Qiyue Wang<sup>2</sup> and ♦Bingyi Jing<sup>3</sup>.* <sup>1</sup>Jiangxi Univ. of Finiance & Economics <sup>2</sup>HKUST <sup>3</sup>SUSTech  
 15:25 Deep Learning for Cell Type Identification Base on Single-Cell Chromatin Accessibility Data  
*Rui Jiang.* 清华大学  
 15:50 A Fast Method for Inference of Phylogenetic Networks  
*Louxin Zhang.* National University of Singapore  
 16:15 Calibrating p Values for Peptide Identification  
*Juntao Zhao<sup>1</sup>, Sheng Lian<sup>1</sup>, Zhen Zhang<sup>2</sup>, Xiaodan Fan<sup>2</sup>, Ning Li<sup>1</sup> and ♦Weichuan Yu<sup>1</sup>.* <sup>1</sup>HKUST <sup>2</sup>CUHK  
 16:40 Floor Discussion.

### Session 23CHI137: Recent Developments in Machine Learning and Causal Inference

Room: BAOSHAN HALL  
 Organizer: Zhonghua Liu, Columbia University.  
 Chair: Zhonghua Liu, Columbia University.

- 15:00 Ensemble Methods for Testing a Global Null  
*Yaowu Liu.* Southwestern University of Finance and Economics  
 15:25 Policy Learning with Asymmetric Utilities  
*Eli Ben-Michael<sup>1</sup>, Kosuke Imai<sup>2</sup> and ♦Zhichao Jiang<sup>3</sup>.* <sup>1</sup>Carnegie Mellon University <sup>2</sup>Harvard University <sup>3</sup>Sun Yat-sen University  
 15:50 Semiparametric Proximal Causal Inference  
 ♦*Yifan Cui<sup>1</sup>, Hongming Pu<sup>2</sup>, Xu Shi<sup>3</sup>, Wang Miao<sup>4</sup> and Eric Tchetgen Tchetgen<sup>2</sup>.* <sup>1</sup>ZJU <sup>2</sup>UPenn <sup>3</sup>UMich <sup>4</sup>PKU

- 16:15 New  $\sqrt{n}$ -Consistent, Numerically Stable Higher Order Influence Function Estimators  
*Lin Liu.* Shanghai Jiao Tong University
- 16:40 Floor Discussion.

### Session 23CHI138: Recent Developments in Applied Probability and Statistics

Room: CHONGMING HALL

Organizer: Lei Huang, Southwest Jiaotong University.  
Chair: Xiao Fang, The Chinese University of Hong Kong.

- 15:00 High-Dimensional Dimension Reduction and Its Application to Classification  
♦*Zhibo Cai<sup>1</sup>, Yingcun Xia<sup>2</sup> and Weiqiang Hang<sup>2</sup>.* <sup>1</sup>Renmin University of China <sup>2</sup>Nantional University of Singapore
- 15:25 From p-Wasserstein Bounds to Moderate Deviations  
♦*Xiao Fang<sup>1</sup> and Yuta Koike<sup>2</sup>.* <sup>1</sup>The Chinese University of Hong Kong <sup>2</sup>University of Tokyo
- 15:50 A Two-Way Heterogeneity Model for Dynamic Networks  
♦*Binyan Jiang<sup>1</sup>, Chenlei Leng<sup>2</sup>, Ting Yan<sup>3</sup>, Qiwei Yao<sup>4</sup> and Xinyang Yu<sup>1</sup>.* <sup>1</sup>The Hong Kong Polytechnic University <sup>2</sup>University of Warwick <sup>3</sup>Central China Normal University <sup>4</sup>London School of Economics
- 16:15 Estimating Conditional Covariance Matrices Dependent on Exogenous Variables  
♦*Hui Jiang and Jinze Ji.* Huazhong University of Science and Technology
- 16:40 Floor Discussion.

### Session 23CHI139: Statistical Modeling for Neuroimaging Data

Room: NANSHI HALL

Organizer: Jing Zhou, Renmin University of China.  
Chair: Jing Zhou, Renmin University of China.

- 15:00 Bayesian Spatially Varying Weight Neural Networks with the Soft-Thresholded Gaussian Process Prior  
♦*Ben Wu<sup>1</sup>, Keru Wu<sup>2</sup> and Jian Kang<sup>3</sup>.* <sup>1</sup>Renmin University of China <sup>2</sup>Duke University <sup>3</sup>University of Michigan
- 15:25 Minimizing Estimated Risks on Unlabeled Data for Semi-Supervised Medical Image Segmentation  
♦*Fuping Wu<sup>1</sup> and Xiahai Zhuang<sup>2</sup>.* <sup>1</sup>University of Oxford <sup>2</sup>Fudan University
- 15:50 Ensembled Seizure Detection Based on Small Training Samples  
♦*Peifeng Tong<sup>1</sup>, Haoxiang Zhan<sup>2</sup> and Songxi Chen<sup>3</sup>.*  
<sup>1</sup>Guanghua School of Management, Peking University, Beijing 100871, China <sup>2</sup>School of Mathematical Science, Peking University, Beijing 100871, China <sup>3</sup>School of Mathematical Science and Guanghua School of Management, Peking University, Beijing 100871, China
- 16:15 A Semiparametric Gaussian Mixture Model for Chest Ct-Based 3d Blood Vessel Reconstruction  
*Qianhan Zeng.* PEKING UNIVERSITY
- 16:40 Floor Discussion.

### Session 23CHI140: Network Data and Large-Scale Computing

Room: PUTUO HALL

Organizer: Wei Lan, Southwestern University of Finance and Economics.  
Chair: Wei Lan, Southwestern University of Finance and Economics.

- 15:00 Subsampling-Based Modified Bayesian Information Criterion for Large-Scale Stochastic Block Models  
♦*Jiayi Deng<sup>1</sup>, Danyang Huang<sup>1</sup>, Xiangyu Chang<sup>2</sup> and Bo Zhang<sup>1</sup>.* <sup>1</sup>Renmin University of China <sup>2</sup>Xi'an Jiaotong University
- 15:25 Statistical Analysis of Fixed Mini-Batch Gradient Descent Estimator  
♦*Haobo Qi<sup>1</sup>, Feifei Wang<sup>2</sup> and Hansheng Wang<sup>1</sup>.* <sup>1</sup>Peking University <sup>2</sup>Renmin University of China
- 15:50 Quasi-Score Matching Estimation for Spatial Autoregressive Model with Random Weights Matrix and Regressors  
♦*Xuan Liang and Tao Zou.* The Australian National University
- 16:15 Identifying Temporal Pathways using Biomarkers in the Presence of Latent Non-Gaussian Components  
♦*Shanghong Xie<sup>1</sup>, Donglin Zeng<sup>2</sup> and Yuanjia Wang<sup>3</sup>.* <sup>1</sup>Southwestern University of Finance and Economics <sup>2</sup>University of North Carolina at Chapel Hill <sup>3</sup>Columbia University
- 16:40 Floor Discussion.

### Session 23CHI170: Checking Structural Change of Complex Data

Room: FENGXIAN HALL

Organizer: Lixing Zhu, Beijing Normal University.  
Chair: Falong Tan, Hunan University.

- 15:00 Score Function-Based Tests for Ultrahigh-Dimensional Linear Models  
♦*Weichao Yang, Xu Guo and Lixing Zhu.* Beijing Normal University
- 15:25 Quantile Regression for Complex Longitudinal Data  
♦*Xuerui Li, Yanyan Liu and Yuanshan Wu.*
- 15:50 Weighted Residual Empirical Processes, Martingale Transformations, and Model Checking for Regressions  
♦*Falong Tan<sup>1</sup>, Xu Guo<sup>2</sup> and Lixing Zhu<sup>3</sup>.* <sup>1</sup>Hunan University <sup>2</sup>Beijing normal university <sup>3</sup>Beijing normal university at Zhuhai
- 16:15 Multiple Change Point Detection in Tensors  
♦*Jiaqi Huang<sup>1</sup>, Junhui Wang<sup>2</sup>, Lixing Zhu<sup>3</sup> and Xuehu Zhu<sup>4</sup>.* <sup>1</sup>Beijing Normal Univeisity <sup>2</sup>Chinese University of Hong Kong <sup>3</sup>Beijing Normal University <sup>4</sup>Xi'an Jiaotong University
- 16:40 Floor Discussion.

### Session 23CHI25: Current Topics in Biostatistics II

Room: LONGFENG HALL-II

Organizer: Somnath Datta, University of Florida.  
Chair: Muxuan Liang, University of Florida.

- 15:00 Inference with Non-Differentiable Surrogate Loss in a General High-Dimensional Classification Framework  
 ♦*Muxuan Liang*<sup>1</sup>, *Yang Ning*<sup>2</sup>, *Maureen Smith*<sup>3</sup> and *Ying Qi Zhao*<sup>4</sup>. <sup>1</sup>University of Florida <sup>2</sup>Cornell University  
<sup>3</sup>University of Wisconsin-Madison <sup>4</sup>Fred Hutchinson Cancer Center
- 15:25 Bayesian Regression for Correlated Compositional Outcomes  
*Nanhua Zhang*. University of Cincinnati
- 15:50 Standardization of Continuous and Categorical Covariates in Sparse Penalized Regressions  
*Xiang Li*<sup>1</sup>, ♦*Qing Pan*<sup>2</sup> and *Yong Ma*<sup>3</sup>. <sup>1</sup>Capital One  
<sup>2</sup>George Washington University <sup>3</sup>US FDA
- 16:15 Floor Discussion.

### Session 23CHI29: Advanced Statistical Considerations on Contemporary Clinical Trials

Room: XUHUI HALL  
 Organizer: Bo Fu, Astellas.  
 Chair: Lin Liu, Astellas.

- 15:00 Advancements and Challenges in Clinical Research Design for Rare Diseases  
*Hou Yan*.
- 15:25 Mendelian Randomization for Drug Target Discovery  
*Xiao Xiang*. Fosun Pharma
- 15:50 Phase II Trial Design Based on Success Probabilities for Phase III in Oncology: a Case Study  
*Xun Liao*. 默克雪兰诺（北京）医药研发有限公司
- 16:15 Floor Discussion.

### Session 23CHI61: Recent Methodological Advances in Survey Statistics

Room: LONGFENG HALL-III  
 Organizer: Cindy Yu, Iowa State University.  
 Chair: Jae-Kwang Kim, Iowa State University.

- 15:00 A Multivariate Bayesian Hierarchical Model for Small Area Estimation of Criminal Victimization Rates in Domains Defined by Age and Gender  
*Emily Berg*. Iowa State University
- 15:25 A-Optimal Split Questionnaire Designs  
*Zhengyuan Zhu*. Iowa State University
- 15:50 Augmented Two-Step Estimating Equations with Nuisance Functionals and Complex Survey Data  
 ♦*Puying Zhao*<sup>1</sup> and *Changbao Wu*<sup>2</sup>. <sup>1</sup>Yunnan University  
<sup>2</sup>University of Waterloo
- 16:15 Floor Discussion.

### Session 23CHI94: Modern Statistical Process Control and Change-Point Problems II

Room: HONGKOU HALL  
 Organizer: Yajun Mei, H. Milton Stewart School of Industrial and Systems Engineering Georgia Institute of Technology; Dongdong Xiang, School of Statistics, East China Normal University.  
 Chair: Liu Liu, College of Mathematics and Physics, Chengdu University of Technology.

- 15:00 Phase II Control Chart Based on Likelihood Ratio Test for Monitoring Time Between Events of Gumbel's Bivariate Exponential Distribution  
 ♦*Jiujun Zhang* and *Peile Chen*.
- 15:25 Low-Rank Matrix Estimation in the Presence of Change-Points  
*Lei Shi*<sup>1</sup>, ♦*Guanghui Wang*<sup>2</sup> and *Changliang Zou*<sup>3</sup>. <sup>1</sup>University of California, Berkeley <sup>2</sup>East China Normal University <sup>3</sup>Nankai University
- 15:50 Identification of Outlying Observations for Large-Dimensional Data  
*Tao Wang*<sup>1</sup>, *Xiaona Yang*<sup>2</sup>, *Yunfei Guo*<sup>3</sup> and ♦*Zhonghua Li*<sup>3</sup>. <sup>1</sup>Huaiyin Normal University <sup>2</sup>Heilongjiang University <sup>3</sup>Nankai University
- 16:15 Dynamic Modeling and Online Monitoring of Tensor Data Streams with Application to Passenger Flow Surveillance  
*Wendong Li*. Shanghai University of Finance and Economics
- 16:40 Floor Discussion.

**July 2, 17:00-18:40**

### Session 23CHI132: Advanced Methods in Adaptive Randomization Designs

Room: JIADING HALL  
 Organizer: Hongjian Zhu, AbbVie Inc..  
 Chair: Wei Ma, Renmin University of China.

- 17:00 Discussion: New Developments in Adaptive Randomization Designs  
*Li-Xin Zhang*. Zhejiang University
- 17:25 Model-Based Adaptive Randomization Procedures for Heteroscedasticity of Treatment Responses  
*Zhongqiang Liu*. Henan Polytechnic University
- 17:50 A New and Unified Family of Covariate Adaptive Randomization Procedures and their Properties  
 ♦*Wei Ma*<sup>1</sup>, *Ping Li*<sup>2</sup>, *Li-Xin Zhang*<sup>3</sup> and *Feifang Hu*<sup>4</sup>. <sup>1</sup>Renmin University of China <sup>2</sup>LinkedIn Corporation <sup>3</sup>Zhejiang University <sup>4</sup>George Washington University
- 18:15 Balancing Unobserved Covariates with Covariate-Adaptive Randomized Experiments  
 ♦*Yang Liu*<sup>1</sup> and *Feifang Hu*<sup>2</sup>. <sup>1</sup>Renmin University of China  
<sup>2</sup>George Washington University
- 18:40 Discussant: Lixin Zhang, Zhejiang University  
 Floor Discussion.

### Session 23CHI14: New Developments in Nonparametric and Semiparametric Methods for Complex Data

Room: XUHUI HALL  
 Organizer: Lan Xue, Oregon State University.  
 Chair: Lan Xue, Oregon State University.

- 17:00 Latent Single-Index Models with Factor Structure for Multivariate Ordinal Data  
*Zhiyong Chen*. School of Mathematics and Statistics, Fujian Normal University

- 17:25 Electricity Consumption Forecasting by a New Neural Network Model: Panel Semiparametric Quantile Regression Neural Network (Psqrnn)  
*Xingcai Zhou, ♦Jiangyan Wang, Hongxia Wang and Jinguan Lin.* Nanjing Audit University
- 17:50 Compositional Multivariate Regression for Microbiome Data Integration  
*Chenxiao Hu, Ying Dai, Thomas Sharpton and ♦Duo Jiang.* Oregon State University
- 18:15 Integrated Subgroup Identification from Multi-Source Data  
*Lihui Shao, ♦Jiaqi Wu, Weiping Zhang and Yu Chen.* University of Science and Technology of China
- 18:40 Floor Discussion.

### Session 23CHI143: Recent Developments of Statistical Methods on Missing Data with Applications

Room: HONGKOU HALL

Organizer: Ming-Hui Chen, University of Connecticut.

Chair: Fang Liu, Soochow University.

- 17:00 A Self-Censoring Model for Multivariate Nonignorable Non-monotone Missing Data  
*♦Yilin Li<sup>1</sup>, Wang Miao<sup>1</sup>, Ilya Shpitser<sup>2</sup> and Eric Tchetgen Tchetgen<sup>3</sup>.* <sup>1</sup>Peking University <sup>2</sup>Johns Hopkins University <sup>3</sup>The Wharton School of the University of Pennsylvania
- 17:25 Bayesian Diagnostics of Hidden Markov Structural Equation Models with Missing Data  
*♦Jingcheng Cai<sup>1</sup>, Ming Ouyang<sup>2</sup>, Kai Kang<sup>1</sup> and Xinyuan Song<sup>3</sup>.* <sup>1</sup>Sun Yat-sen University <sup>2</sup>The Chinese University of Hong Kong <sup>3</sup>The Chinese University of Hong Kong,
- 17:50 Bayesian Modeling and Inference for Item Response Model with Nonignorable Missing Data  
*♦Zhihua Ma<sup>1</sup>, Jing Wu<sup>2</sup> and Ming-Hui Chen<sup>3</sup>.* <sup>1</sup>Shenzhen University <sup>2</sup>The University of Rhode Island <sup>3</sup>University of Connecticut
- 18:15 Floor Discussion.

### Session 23CHI157: Statistical Advances in Biomedical Data Analysis

Room: NANSHI HALL

Organizer: Lin Hou, Center for Statistical Science, Tsinghua University.

Chair: Huaying Fang, Academy for Multidisciplinary Studies, Capital Normal University.

- 17:00 A Hybrid Machine Learning and Regression Method for Cell Type Deconvolution of Spatial Barcoding-Based Transcriptomic Data  
*Yunqing Liu<sup>1</sup>, ♦Ningshan Li<sup>2</sup>, Ji Qi<sup>3</sup>, Gang Xu<sup>4</sup>, Jiayi Zhao<sup>3</sup>, Nating Wang<sup>3</sup>, Aurélien Juster<sup>5</sup>, Taylor Adams<sup>5</sup>, Zuoheng Wang<sup>3</sup> and Xiting Yan<sup>5</sup>.* <sup>1</sup>Department of Biostatistics, Yale School of Public Health <sup>2</sup>The Second Affiliated Hospital, School of Medicine, The Chinese University of Hong Kong , Shenzhen & Longgang District People's Hospital of Shenzhen <sup>3</sup>Department of Biostatistics, Yale School of Public Health <sup>4</sup>Department of Mathematical Sciences, University of Nevada <sup>5</sup>Section of Pulmonary, Critical Care and Sleep Medicine, Yale School of Medicine

- 17:25 Dimensionality Reduction with Network Regularization in Single-Cell Expression Analysis  
*Huaying Fang.* Academy for Multidisciplinary Studies, Capital Normal University
- 17:50 Time-Varying Treatment Effects of Functional Data with Latent Confounders: Application to Sleep Heart Health Studie  
*♦Jie Li<sup>1</sup>, Shujie Ma<sup>2</sup> and Yehua Li<sup>2</sup>.* <sup>1</sup>Renmin University of China <sup>2</sup>University of California, Riverside
- 18:15 A Unified Quantile Framework Reveals Nonlinear Heterogeneous Transcriptome-Wide Associations  
*♦Tianying Wang<sup>1</sup>, Iuliana Ionita-Laza<sup>2</sup> and Ying Wei<sup>2</sup>.*  
<sup>1</sup>Tsinghua University <sup>2</sup>Columbia University
- 18:40 Floor Discussion.

### Session 23CHI158: Statistical Inferences in Computational Biology and Genetics

Room: PUTUO HALL

Organizer: Qizhai Li, Academy of Mathematics and Systems Science, CAS.

Chair: Qizhai Li, Academy of Mathematics and Systems Science, CAS.

- 17:00 A Data-Adaptive Bayesian Regression Approach for Polygenic Risk Prediction  
*Lin Hou.* Tsinghua University
- 17:25 Fastancom: a Fast Method for Analysis of Compositions of Microbiomes  
*Tao Wang.* Shanghai Jiao Tong University
- 17:50 Feature Screening for Clustering Analysis with Applications Single-Cell Rna Sequencing  
*Changhu Wang, Zihao Chen and ♦Ruixin Xi.* Peking University
- 18:15 Robust and Powerful Gene-Environment Interaction Tests using Rare Genetic Variants in Case-Control Studies  
*♦Yanan Zhao and Hong Zhang.*
- 18:40 Floor Discussion.

### Session 23CHI18: Advanced Statistical Methods for Complex Data

Room: CHANGNING HALL

Organizer: Linlin Dai, Southwestern University of Finance and Economics; Gang Li, UCLA.

Chair: Linlin Dai, Southwestern University of Finance and Economics.

- 17:00 Moment Deviation Subspaces of Dimension Reduction for High-Dimensional Data with Change Structure  
*♦Xuehu Zhu<sup>1</sup>, Luoyao Yu<sup>1</sup>, Jiaqi Huang<sup>2</sup>, Junmin Liu<sup>1</sup> and Lixing Zhu<sup>2</sup>.* <sup>1</sup>Xi'an Jiaotong University <sup>2</sup>Beijing Normal University
- 17:25 Health Utility Survival for Randomized Clinical Trials: a Composite Endpoint for Clinical Trial Designs  
*Yangqing Deng<sup>1</sup>, ♦Meiling Hao<sup>2</sup>, John R. De Almeida<sup>3</sup>, Xiaolu Wang<sup>2</sup> and Wei Xu<sup>4</sup>.* <sup>1</sup>Princess Margaret Cancer Centre <sup>2</sup>University of International Business and Economics <sup>3</sup>University Health Network <sup>4</sup>Princess Margaret Cancer Centre; University of Toronto

- 17:50 Statistical Inference in High-Dimensional Regression with Streaming Data  
 ♦*Ruijian Han<sup>1</sup>, Lan Luo<sup>2</sup>, Yuanyuan Lin<sup>3</sup> and Jian Huang<sup>1</sup>*.  
<sup>1</sup>The Hong Kong Polytechnic University <sup>2</sup>University of Iowa  
<sup>3</sup>The Chinese University of Hong Kong
- 18:15 Quantile Autoregressive Conditional Heteroscedasticity  
 ♦*Qianqian Zhu<sup>1</sup>, Songhua Tan<sup>1</sup>, Yao Zheng<sup>2</sup> and Guodong Li<sup>3</sup>*. <sup>1</sup>Shanghai University of Finance and Economics <sup>2</sup>University of Connecticut <sup>3</sup>University of Hong Kong
- 18:40 Floor Discussion.

### Session 23CHI19: Recent Development in Extreme Value Theory

Room: JINGAN HALL  
 Organizer: Liuhua Peng, University of Melbourne.  
 Chair: Liuhua Peng, University of Melbourne.

- 17:00 Estimation and Inference for Extreme Continuous Treatment Effects  
*Wei Huang<sup>1</sup>, Shuo Li<sup>2</sup> and Liuhua Peng<sup>1</sup>*. <sup>1</sup>The University of Melbourne <sup>2</sup>Tianjin University of Finance and Economics
- 17:25 Panel Quantile Regression for Extreme Risk  
*Yanxi Hou<sup>1</sup>, Xuan Leng<sup>2</sup>, Liang Peng<sup>3</sup> and Yinggang Zhou<sup>2</sup>*. <sup>1</sup>Fudan University <sup>2</sup>Xiamen University <sup>3</sup>Georgia State University
- 17:50 Simultaneous Confidence Bands for Conditional Value-at-Risk and Expected Shortfall  
*Shuo Li<sup>1</sup>, Liuhua Peng<sup>2</sup> and Xiaojun Song<sup>3</sup>*. <sup>1</sup>Tianjin University of Finance and Economics <sup>2</sup>The University of Melbourne <sup>3</sup>Peking University
- 18:15 Floor Discussion.

### Session 23CHI20: High Dimensional Modeling and Inference

Room: NANHUI HALL  
 Organizer: Jingyuan Liu, Xiamen University.  
 Chair: Jingyuan Liu, Xiamen University.

- 17:00 Structure Learning of a Gaussian Amp Chain Graph via Acyclicity Constraints  
 ♦*Wei Zhou and Wei Zhong*. Xiamen University
- 17:25 New Tests for High-Dimensional Two-Sample Mean Problems with Consideration of Correlation Structure  
 ♦*Songshan Yang<sup>1</sup>, Shurong Zheng<sup>2</sup> and Runze Li<sup>3</sup>*.  
<sup>1</sup>Renmin University of China <sup>2</sup>Northeast Normal University  
<sup>3</sup>The Pennsylvania State University
- 17:50 Projective Independence Tests in High Dimensions: The Curses and the Cures  
 ♦*Yaowu Zhang<sup>1</sup> and Liping Zhu<sup>2</sup>*. <sup>1</sup>Shanghai University of Finance and Economics <sup>2</sup>Renmin University of China
- 18:15 Large-Scale Two-Sample Comparison of Support Sets  
*Haoyu Geng<sup>1</sup>, Xiaolong Cui<sup>1</sup>, Haojie Ren<sup>2</sup> and Changliang Zou<sup>1</sup>*. <sup>1</sup>Nankai University <sup>2</sup>Shanghai Jiao Tong University
- 18:40 Floor Discussion.

### Session 23CHI21: Statistical Learning for Regression Analysis and Change Point Detection

Room: SONGJIANG HALL  
 Organizer: Wei Zhong, Xiamen University.  
 Chair: Wei Zhong, Xiamen University.

- 17:00 Online Smooth Backfitting for Generalized Additive Models  
 ♦*Ying Yang<sup>1</sup>, Fang Yao<sup>2</sup> and Peng Zhao<sup>3</sup>*. <sup>1</sup>Academy of Mathematics and Systems Science, Chinese Academy of Sciences <sup>2</sup>Peking University <sup>3</sup>Jiangsu Normal University
- 17:25 Composite Smoothed Quantile Regression  
*Xiaozhou Wang*. East China Normal University
- 17:50 Detecting Multi-Threshold Effects Ofaccelerate Failure Time Model by Sample-Splitting Methods  
*Chuang Wan*.
- 18:15 Discussant: Chuang Wan, Nankai University
- 18:40 Floor Discussion.

### Session 23CHI23: Statistics Process Control and Change-Point Analysis

Room: JINSHAN HALL  
 Organizer: Dong Han, Department of Statistics, School of Mathematical Sciences, Shanghai Jiao Tong University.  
 Chair: Dong Han, Department of Statistics, School of Mathematical Sciences, Shanghai Jiao Tong University.

- 17:00 Stochastic Integral Bootstrap for Statistics of Irregularly Spaced Spatial Data  
*Shibin Zhang*. Shanghai Normal University
- 17:25 An Optimization Method for Monitoring Change-Point in Small Samples  
 ♦*Dong Han<sup>1</sup>, Fugee Tsung<sup>2</sup> and Jinguo Xian<sup>1</sup>*. <sup>1</sup>Shanghai Jiao Tong University <sup>2</sup>Hong Kong University of Science and Technology
- 17:50 Surface Temperature Monitoring in Liver Procurement via Functional Variance Change-Point Analysis  
 ♦*Zhenguo Gao<sup>1</sup>, Pang Du<sup>2</sup>, Ran Jin<sup>2</sup> and John Robberson<sup>2</sup>*.  
<sup>1</sup>Shanghai Jiao Tong University <sup>2</sup>Virginia Tech
- 18:15 Robust Online Detection in Serially Correlated Directed Network  
 ♦*Miaomiao Yu<sup>1</sup>, Yuhao Zhou<sup>2</sup> and Fugee Tsung<sup>3</sup>*. <sup>1</sup>East China Normal University <sup>2</sup>University of North Carolina at Chapel Hill <sup>3</sup>Hong Kong University of Science and Technology
- 18:40 Floor Discussion.

### Session 23CHI24: Recent Developments in the Analysis of Complex Lifetime Data

Room: BAOSHAN HALL  
 Organizer: Hua Shen, University of Calgary.  
 Chair: Hua Shen, University of Calgary.

- 17:00 Differential Equation-Assisted Local Polynomial Regression  
*W. John Braun*. UBC
- 17:25 Covariate Balancing with Measurement Error  
*Xialing Wen and Ying Yan*. Sun Yat-sen University

- 17:50 Variable Selection for Recurrent Event Model with Covariates Subject to Measurement Error  
 ♦Kaida Cai<sup>1</sup>, Hua Shen<sup>2</sup> and Xuewen Lu<sup>2</sup>. <sup>1</sup>Southeast University <sup>2</sup>University of Calgary

- 18:15 Chair and Discussant  
*Hua Shen.* University of Calgary  
 18:40 Discussant: Hua Shen, University of Calgary  
 Floor Discussion.

### **Session 23CHI27: Stochastic Modeling and Inference of Epidemiological and Industrial Data with Complex Structures**

Room: CHONGMING HALL  
 Organizer: Liqun Wang, University of Manitoba.  
 Chair: Liqun Wang, University of Manitoba.

- 17:00 Statistical Inference of Multi-State Transition Model for Longitudinal Data with Measurement Error and Heterogeneity  
 ♦Jing Guan and Jiajie Qin. Tianjin University
- 17:25 Bi-Level Selection with the Sparse Group Bar for Multivariate Interval- Censored Data  
 ♦Xuewen Lu and Fatemeh Mahmoudi. University of Calgary
- 17:50 Euler's Number, Euler Numbers, and Eulerian Numbers  
*James C. Fu<sup>1</sup>, Wan-Chen Lee and ♦Hsing-Ming Chang<sup>2</sup>.*  
<sup>1</sup>University of Manitoba <sup>2</sup>National Cheng Kung University
- 18:15 A Censored Quantile Transformation Model for Alzheimer's Disease Data with Multiple Functional Covariates  
*Maozai Tian.* Renmin University of China
- 18:40 Floor Discussion.

### **Session 23CHI79: Novel Bayesian Methods for Complex Dataand their Applications**

Room: T-1  
 Organizer: Yichuan Zhao, Georgia State University.  
 Chair: Yichuan Zhao, Georgia State University.

- 17:00 A Bayesian Joint Model of Longitudinal and Time-to-Event Data for the Aids Treatment Effectiveness Evaluation  
 ♦Tao Wang and Yinxiang Zhang.
- 17:25 Statistical Inference for Copula-Based Dependent Competing Risks Model with Step-Stress Accelerated Life Test  
*Wenhao Gui.* Beijing Jiaotong University
- 17:50 Bayesian Jackknife Empirical Likelihood for the Error Variance in Linear Regression Models  
 ♦Hongyan Jiang<sup>1</sup> and Yichuan Zhao<sup>2</sup>. <sup>1</sup>Department of Mathematics and Physics, Huaiyin Institute of Technology, Huai'an, People's Republic of China <sup>2</sup>Department of Mathematics and Statistics, Georgia State University, Atlanta, GA, United States
- 18:15 Changepoint Joint Modeling of High-Dimensional Longitudinal and Survival Data: a Bayesian Approach  
*Yanxin Huang.* University of South Florida
- 18:40 Floor Discussion.

### **July 3, 8:30-10:10**

### **Session 23CHI1: Sufficient Dimension Reduction and Beyond**

Room: YANGPU HALL  
 Organizer: Yingcun Xia, National University of Singapore, University of Electronic Science and Technology.  
 Chair: Yingcun Xia, National University of Singapore, University of Electronic Science and Technology.

- 8:30 Deep Nonlinear Sufficient Dimension Reduction  
*Fengyin Chen<sup>1</sup>, Yuling Jiao<sup>2</sup>, Rui Qiu<sup>1</sup> and ♦Zhou Yu<sup>1</sup>.*  
<sup>1</sup>East China Normal University <sup>2</sup>Wuhan University
- 8:55 Dimension Reduction for Frechet Regression  
*Qi Zhang, Lingzhou Xue and ♦Bing Li.* Pennsylvania State University
- 9:20 Sliced Inverse Regression with Large Structural Dimension  
*Dongming Huang<sup>1</sup>, Songtao Tian<sup>2</sup> and ♦Qian Lin<sup>2</sup>.* <sup>1</sup>NUS <sup>2</sup>Tsinghua University
- 9:45 A Unified Generalization of Inverse Regression via Adaptive Column Selection  
*Yin Jin and ♦Wei Luo.* Zhejiang University
- 10:10 Floor Discussion.

### **Session 23CHI10: Novel Bayesian Adaptive Clinical Trial Designs and Methods for Precision Medicine**

Room: JINGAN HALL  
 Organizer: Ying Yuan, University of Texas MD Anderson Cancer Center.  
 Chair: Mengyi Lu, Nanjing Medical University.

- 8:30 Sam: Self-Adapting Mixture Prior to Dynamically Borrow Information from Historical Data in Clinical Trials  
*Peng Yang<sup>1</sup>, Yuansong Zhao<sup>2</sup>, Lei Nie<sup>3</sup>, Jonathon Vallejo<sup>3</sup> and ♦Ying Yuan<sup>4</sup>.* <sup>1</sup>Rice University <sup>2</sup>The University of Texas Health Science Center <sup>3</sup>FDA <sup>4</sup>The University of Texas MD Anderson Cancer Center
- 8:55 Adaptive Promising Zone Design for Cancer Immunotherapy with Heterogenous Delayed Treatment Effects  
*Bosheng Li<sup>1</sup>, Fangrong Yan<sup>1</sup> and ♦Depeng Jiang<sup>2</sup>.* <sup>1</sup>China Pharmaceutical University <sup>2</sup>University of Manitoba
- 9:20 A Generalized Phase 1-2-3 Design Integrating Dose Optimization with Confirmatory Treatment Comparison  
*Yong Zang.* Indiana University
- 9:45 Discussant: Xiaoping Su, University of Texas MD Anderson Cancer Center
- 10:10 Floor Discussion.

### **Session 23CHI11: Advanced Biostatistics and Bioinformatics Approaches for Medical Research**

Room: NANHUI HALL  
 Organizer: Zhengjia Chen, University of Illinois at Chicago.  
 Chair: Yue Cui, Missouri State University.

- 8:30 The Identifiability of Copula Models for Dependent Competing Risks Data with Exponentially Distributed Margins  
*Antai Wang.* New Jersey Institute of Technology

- 8:55 Adaptive Bayesian Phase i Clinical Trial Designs for Estimating the Maximum Tolerated Doses for Two Drugs While Fully Utilizing all Toxicity Information  
*Zhengjia Chen.* University of Illinois at Chicago
- 9:20 Developing Machine Learning Models to Improve Fracture Prediction using Genomic and Phenotypic Data in Large-Scale Observational Studies  
*Qing Wu.* The Ohio State University
- 9:45 using Auxiliary Information in Probability Survey Data to Improve Pseudo-Weighting in Non-Probability Samples: a Copula Model Approach  
*Tingyu Zhu, ♦Lan Xue and Ginny Lesser.* Oregon State University
- 10:10 Floor Discussion.

### Session 23CHI12: Topics on Latent Variable Models and Minorisation-Maximisation Algorithm

Room: SONGJIANG HALL  
Organizer: Dalei Yu, Xian Jiaotong University.  
Chair: Dalei Yu, Xian Jiaotong University.

- 8:30 Sufficient Dimension Reductions under the Mixtures of Multivariate Elliptical Distributions  
*Wenjuan Li, Hongming Pei, Ali Jiang and ♦Fei Chen.* Yunnan University of Finance and Economics
- 8:55 Earthquake Parametric Insurance with Bayesian Spatial Quantile Regression  
*Jeffrey Pai<sup>1</sup>, ♦Yunxian Li<sup>2</sup>, Aijun Yang<sup>3</sup> and Chenxu Li<sup>2</sup>.*  
<sup>1</sup>University of Manitoba <sup>2</sup>Yunnan University of Finance and Economics <sup>3</sup>Nanjing Forest University
- 9:20 Proportional Inverse Gaussian Distribution: a New Tool for Analysing Continuous Proportional Data  
*♦Pengyi Liu<sup>1</sup>, Guo-Liang Tian<sup>2</sup>, Kam Chuen Yuen<sup>3</sup>, Chi Zhang<sup>4</sup> and Man-Lai Tang<sup>5</sup>.* <sup>1</sup>Yunnan University of Finance and Economics <sup>2</sup>Southern University of Science and Technology <sup>3</sup>The University of Hong Kong <sup>4</sup>Shenzhen University <sup>5</sup>Brunel University London
- 9:45 Floor Discussion.

### Session 23CHI13: New Developments in Statistical Detection

Room: JINSHAN HALL  
Organizer: Maoyuan Zhou, Civil Aviation University of China.  
Chair: Maoyuan Zhou, Civil Aviation University of China.

- 8:30 Modeling Autoregressive Conditional Regional Extremes with Application to Solar Flare Detection  
*Jili Wang<sup>1</sup> and ♦Zhengjun Zhang<sup>2</sup>.* <sup>1</sup>University of Wisconsin <sup>2</sup>University of Chinese Academy of Sciences
- 8:55 A Fdr – based Method for Monitoring High Dimensional Data Streams  
*♦Dequan Qi and Nan Liang.* School of Mathematics and Statistics, Changchun University of Science and Technology, Changchun Jilin

- 9:20 Two Robust Multivariate Exponentially Weighted Moving Average Charts to Facilitate Distinctive Product Quality Features Assessment  
*♦Zhi Song<sup>1</sup>, Amitava Mukherjee<sup>2</sup>, Peihua Qiu<sup>3</sup> and Maoyuan Zhou<sup>4</sup>.* <sup>1</sup>Shenyang Agricultural University <sup>2</sup>XLRI-Xavier School of Management <sup>3</sup>University of Florida <sup>4</sup>Civil Aviation University of China
- 9:45 Enhancing Modeling and Monitoring Strategy for Directed Count-Weighted Networks  
*♦Chunjie Wu, Jinhua Qin and Wendong Li.* Shanghai University of Finance and Economics
- 10:10 Floor Discussion.

### Session 23CHI141: Factor Modelling and Data Analysis in Large-Scale Datasets

Room: BAOSHAN HALL  
Organizer: Danyang Huang, Renmin University of China.  
Chair: Danyang Huang, Renmin University of China.

- 8:30 Factor Modelling for High-Dimensional Functional Time Series  
*Shaojun Guo.* Renmin University of China
- 8:55 Network Gradient Descent Algorithm for Decentralized Federated Learning  
*♦Shuyuan Wu<sup>1</sup>, Danyang Huang<sup>2</sup> and Hansheng Wang<sup>1</sup>.*  
<sup>1</sup>Peking University <sup>2</sup>Renmin University of China
- 9:20 Wasserstein Distance-Based Spectral Clustering with Application to Transaction Data  
*♦Yingqiu Zhu<sup>1</sup>, Danyang Huang<sup>2</sup> and Bo Zhang<sup>2</sup>.*  
<sup>1</sup>University of International Business and Economics <sup>2</sup>Renmin University of China
- 9:45 High-Dimensional Inference for Dynamic Treatment Effects  
*Yuqian Zhang.*
- 10:10 Floor Discussion.

### Session 23CHI142: Recent Advances in Spatio Temporal Modelling

Room: CHONGMING HALL  
Organizer: Yingying Ma, Beihang University.  
Chair: Yingying Ma, Beihang University.

- 8:30 On Semiparametrically Dynamic Functional-Coefficient Autoregressive Spatio-Temporal Models with Irregular Location Wide Nonstationarity  
*Rongmao Zhang.* Zhejiang University
- 8:55 One-Way or Two-Way Factor Model for Matrix Sequences?  
*Yong He<sup>1</sup>, ♦Xinbing Kong<sup>2</sup>, Lorenzo Trapani<sup>3</sup> and Long Yu<sup>4</sup>.* <sup>1</sup>Shandong University <sup>2</sup>Nanjing Audit University <sup>3</sup>Nottingham University <sup>4</sup>Shanghai University of Finance and Economics
- 9:20 Multivariate Reduced-Rank Spatiotemporal Models  
*Dan Pu<sup>1</sup>, Kuangnan Fang<sup>1</sup>, ♦Wei Lan<sup>2</sup> and Qingzhao Zhang<sup>1</sup>.* <sup>1</sup>Xiamen University <sup>2</sup>Southwestern University of Finance and Economics
- 9:45 Design-Based Covariate Adjustment for Causal Inference with Interference and Noncompliance  
*Hanzhong Liu.* Tsinghua University
- 10:10 Floor Discussion.

**Session 23CHI145: Human-Centric Statistical Learning**

Room: LONGFENG HALL-I

Organizer: Xiangyu Chang, Xi'an Jiaotong University.

Chair: Xiangyu Chang, Xi'an Jiaotong University.

- 8:30 Privacy-Preserving Community Detection for Locally Distributed Multiple Networks

♦*Xiao Guo*<sup>1</sup>, *Xiang Li*<sup>2</sup>, *Xiangyu Chang*<sup>3</sup> and *Shujie Ma*<sup>4</sup>.

<sup>1</sup>Northwest University, China <sup>2</sup>Peking University, China

<sup>3</sup>Xi'an Jiaotong University, China <sup>4</sup>University of California-Riverside, USA

- 8:55 Learning Multitask Gaussian Bayesian Networks

♦*Shuai Liu*<sup>1</sup>, *Yixuan Qiu*<sup>2</sup>, *Baojuan Li*<sup>3</sup>, *Huaning Wang*<sup>3</sup> and *Xiangyu Chang*<sup>1</sup>.

<sup>1</sup>Xi'an Jiaotong University <sup>2</sup>Shanghai University of Finance and Economics <sup>3</sup>Air Force Medical University,

- 9:20 2d-Shapley: a Framework for Fragmented Data Valuation

♦*Zhihong Liu*<sup>1</sup>, *Hoang Anh Just*<sup>2</sup>, *Xiangyu Chang*<sup>1</sup>, *Xi Chen*<sup>3</sup> and *Ruoxi Jia*<sup>2</sup>.

<sup>1</sup>Xi'an Jiaotong University <sup>2</sup>Virginia Tech <sup>3</sup>New York University

- 9:45 2d-Shapley: a Framework for Fragmented Data Valuation

♦*Zhihong Liu*<sup>1</sup>, *Hoang Just*<sup>2</sup>, *Xiangyu Chang*<sup>1</sup> and *Ruoxi Jia*<sup>3</sup>.

<sup>1</sup>Xi'an Jiaotong University <sup>2</sup>Virginia Polytechnic Institute and State University <sup>3</sup>Virginia Tech

- 10:10 Toward a Fairness-Aware Scoring System for Algorithmic Decision-Making

♦*Yi Yang*<sup>1</sup>, *Ying Wu*<sup>2</sup>, *Xiangyu Chang*<sup>2</sup>, *Mei Li*<sup>3</sup> and *Yong Tan*<sup>4</sup>.

<sup>1</sup>Xi'an Jiaotong-liverpool University <sup>2</sup>Xi'an Jiaotong University <sup>3</sup>University of Oklahoma <sup>4</sup>University of Washington

Floor Discussion.

10:10 Floor Discussion.

**Session 23CHI148: New Developments on Design of Experiments and Sampling Methods**

Room: FENGXIAN HALL

Organizer: Fasheng Sun, Northeast Normal University.

Chair: Fasheng Sun, Northeast Normal University.

- 8:30 Adaptive Order-of-Addition Experiments via the Quick-Sort Algorithm

♦*Dennis K. J. Lin*<sup>1</sup> and ♦*Jianbin Chen*<sup>2</sup>. <sup>1</sup>Purdue University

<sup>2</sup>Beijing Institute of Technology

- 8:55 Interaction Effects in Pairwise Ordering Model

♦*Chunyan Wang*<sup>1</sup> and *Dennis Lin*<sup>2</sup>. <sup>1</sup>aCenter for Applied Statistics and School of Statistics, Renmin University of China <sup>2</sup>bDepartment of Statistics, Purdue University, West Lafayette

- 9:20 Subsampling Markov Chain Monte Carlo Algorithm Based on Energy Distance

♦*Sumin Wang*<sup>1</sup>, *Fasheng Sun*<sup>2</sup> and *Min-Qian Liu*<sup>1</sup>. <sup>1</sup>Nankai University <sup>2</sup>Northeast Normal University

- 9:45 Construction of Orthogonal Maximin Distance Designs

♦*Wenlong Li*<sup>1</sup>, *Yubin Tian*<sup>1</sup> and *Min-Qian Liu*<sup>2</sup>. <sup>1</sup>Beijing Institute of Technology <sup>2</sup>Nankai University

10:10 Floor Discussion.

**Session 23CHI162: New Methods and Applications of Reinforcement Learning for Complex Data**

Room: LONGFENG HALL-II

Organizer: Xingdong Feng, Shanghai University of Finance and Economics, China.

Chair: .

- 8:30 Value Enhancement of Reinforcement Learning via Efficient and Robust Trust Region Optimization

*Chengchun Shi*<sup>1</sup>, *Zhenglin Qi*<sup>1</sup>, ♦*Jianing Wang*<sup>2</sup> and *Fan Zhou*<sup>1</sup>. <sup>1</sup>Advisor <sup>2</sup>me

- 8:55 Doubly Inhomogeneous Reinforcement Learning

♦*Liyuan Hu*<sup>1</sup>, *Mengbing Li*<sup>2</sup>, *Chengchun Shi*<sup>1</sup>, *Zhenke Wu*<sup>2</sup> and *Piotr Fryzlewicz*<sup>1</sup>. <sup>1</sup>London School of Economics and Political Science <sup>2</sup>University of Michigan, Ann Arbor

- 9:20 Damped Anderson Mixing for Deep Reinforcement Learning: Acceleration, Convergence, and Stabilization

*Ke Sun*<sup>1</sup>, *Yafei Wang*<sup>1</sup>, *Yi Liu*<sup>1</sup>, *Bo Pan*<sup>1</sup>, *Yingnan Zhao*<sup>2</sup>, *Shangling Ju*<sup>3</sup>, *Bei Jiang*<sup>1</sup> and ♦*Linglong Kong*<sup>1</sup>.

<sup>1</sup>University of Alberta <sup>2</sup>Harbin Institute of Technology

<sup>3</sup>Huawei Technologies Ltd

- 9:45 Application of Distributional Reinforcement Learning in Ride-Sharing Industry

*Fan Zhou*.

10:10 Floor Discussion.

**Session 23CHI164: Complex Statistical Modelling and Testing**

Room: LONGFENG HALL-III

Organizer: Xuening Zhu, Fudan University.

Chair: Yimeng Ren, Fudan University.

- 8:30 Trajectory Representation Learning with Multilevel Attention for Driver Identification

♦*Mengyuan Li*<sup>1</sup>, *Yuanyuan Zhang*<sup>2</sup>, *Yaya Zhao*<sup>1</sup>, *Yalei Du*<sup>2</sup> and *Xiaoling Lu*<sup>1</sup>. <sup>1</sup>Renmin University of China <sup>2</sup>Beijing Baixingkefu Network Technology Co., Ltd.

- 8:55 "this Crime is not that Crime"—Classification and Evaluation of Four Common Crimes

♦*Ke Xu*<sup>1</sup>, *Hangyu Liu*<sup>1</sup>, *Fang Wang*<sup>2</sup> and *Hansheng Wang*<sup>3</sup>. <sup>1</sup>University of International Business and Economics <sup>2</sup>Shandong University <sup>3</sup>Peking University

- 9:20 A Geometrical Model with Stochastic Error for Abnormal Motion Detection of Portal Crane Bucket Grab

♦*Baichen Yu*<sup>1</sup>, *Xiao Wang*<sup>2</sup> and *Hansheng Wang*<sup>3</sup>. <sup>1</sup>East China Normal University <sup>2</sup>Qingdao University <sup>3</sup>Peking University

- 9:45 Road Network Enhanced Human Activity Recognition with Spatial-Temporal Transformer

*Yaya Zhao*. Renmin University of China

- 8:30 Test of the Latent Dimension of a Spatial Blind Source Separation Model  
*Christoph Muehlmann, François Bachoc, Klaus Nordhausen and \*Mengxi Yi.*
- 8:55 Testing Sufficiency for Transfer Learning  
♦*Ziqian Lin<sup>1</sup>, Yuan Gao, Feifei Wang<sup>2</sup> and Hansheng Wang.*  
<sup>1</sup>linziqian@stu.pku.edu.cn <sup>2</sup>feifei.wang@ruc.edu.cn
- 9:20 Consistent Selection of the Number of Groups in Panel Models via Sample-Splitting  
♦*Zhe Li<sup>1</sup>, Xuening Zhu and Changliang Zou.* <sup>1</sup>Fudan University
- 9:45 Distributed Estimation and Inference for Spatial Autoregression Model with Large Scale Networks  
♦*Yimeng Ren, Zhe Li, Yuan Gao and Hansheng Wang.*
- 10:10 Floor Discussion.

### Session 23CHI168: Recent Advances in Sequencing Data Analysis, Reinforcement Learning and Missing Data Handling

Room: JIADING HALL

Organizer: Dongmei Li, University of Rochester, Rochester, NY, USA.

Chair: Dongmei Li, University of Rochester, Rochester, NY, USA.

- 8:30 Statistical Methods for Allele-Specific Expression Analysis using Single-Cell Rna-Seq Data  
*Rui Xiao.* University of Pennsylvania

- 8:55 Learning to Make Adherence-Aware Recommendations  
♦*Guanting Chen<sup>1</sup>, Xiaocheng Li<sup>2</sup>, Chunlin Sun<sup>3</sup> and Hanzhao Wang<sup>2</sup>.* <sup>1</sup>University of North Carolina at Chapel Hill <sup>2</sup>Imperial College Business School <sup>3</sup>Stanford University

- 9:20 Rna Sequencing Differential Analysis using Deep-Learning Algorithms  
*Shijian Deng, Zedian Xie and \*Dongmei Li.* University of Rochester

- 9:45 Leveraging Real World Evidence in Regulatory Submissions and Handling Missing Outcome in Real World Data  
♦*Jingyuan Yang, Terry Boodhoo and Hongyan Qiao.* Abb-Vie

10:10 Floor Discussion.

### Session 23CHI3: Computational Approaches to Single-Cell Genomics and Spatial-Omics Data Analysis and Clinical Applications

Room: XUHUI HALL

Organizer: Lana Garmire, University of Michigan.

Chair: Lana Garmire, University of Michigan, USA.

- 8:30 Large-Scale Single-Cell-Based Deconvolution of Pan-Liver Diseases and Spatial Transcriptomics Identified a Novel Cell-Based Marker in Liver Cancer  
*Bin Chen.* Michigan State University

- 8:55 Fastmix: Making Cell Type-Specific Biomarker Inference with Flowcytometry Data and Gene Expressions without Deconvolution  
*Yun Zhang<sup>1</sup>, Hao Sun<sup>2</sup>, Aishwarya Mandava<sup>3</sup>, Brian Aevemann<sup>3</sup>, Tobias Kollmann<sup>4</sup>, Richard Scheuermann<sup>3</sup>, \*Xing*

*Qiu<sup>2</sup> and Yu Qian<sup>5</sup>.* <sup>1</sup>J. Craig Venter Institute <sup>2</sup>University of Rochester <sup>3</sup>J. Craig Venter Institute <sup>4</sup>University of Western Australia <sup>5</sup>J. Craig Venter Institute,

- 9:20 Asgard is a Single-Cell Guided Pipeline to Aid Repurposing of Drugs  
*Bing He, Yao Xiao, Haodong Liang, Qianhui Huang, Yuheng Du, Yijun Li, David Garmire, Duxin Sun and \*Lana Garmire.*
- 9:45 Dissection of the Cell Interaction Landscapes Based on Single-Cell Spatial Transcriptome Data with Artificial Intelligence  
*Runze Li and \*Xuerui Yang.* Tsinghua University
- 10:10 Floor Discussion.

### Session 23CHI43: Recent Advancements in Bayesian Methods for Causal Inference

Room: NANSHI HALL

Organizer: Yisheng Li, University of Texas MD Anderson Cancer Center.

Chair: Yisheng Li, University of Texas MD Anderson Cancer Center.

- 8:30 A Bayesian Non-Parametric Approach for Causal Mediation with a Post-Treatment Confounder  
♦*Woojung Bae, Michael Daniels and Michael Perri.* University of Florida
- 8:55 In Nonparametric and High-Dimensional Models, Bayesian Ignorability is an Informative Prior  
*Antonio Linero.* University of Texas at Austin
- 9:20 Bayesian Semiparametric Models for Dynamic Treatment Strategies with Incomplete Covariate Information  
*Arman Organisian.* Brown University
- 9:45 A Bayesian Machine Learning Approach for Estimating Heterogeneous Survivor Causal Effects: Applications to a Critical Care Trial  
*Xinyuan Chen<sup>1</sup>, Michael Harhay<sup>2</sup>, \*Guangyu Tong<sup>3</sup> and Fan Li<sup>3</sup>.* <sup>1</sup>Mississippi State University <sup>2</sup>University of Pennsylvania <sup>3</sup>Yale University

10:10 Floor Discussion.

### Session 23CHI89: Modern Learning Methods for Causal Inference and Survey Inference

Room: HONGKOU HALL

Organizer: Xinyi Li, Clemson University.

Chair: Yifan Cui, Zhejiang University.

- 8:30 Inference for Treatment Effects on Multiple Derived Outcomes  
♦*Yumou Qiu<sup>1</sup>, Jiarui Sun<sup>2</sup> and Xiao-Hua Zhou<sup>2</sup>.* <sup>1</sup>Iowa State University <sup>2</sup>Peking University
- 8:55 Probability Weighted Clustered Coefficients Regression Models in Complex Survey Sampling  
♦*Mingjun Gang<sup>1</sup>, Xin Wang<sup>2</sup>, Zhonglei Wang<sup>1</sup> and Wei Zhong<sup>1</sup>.* <sup>1</sup>Xiamen University <sup>2</sup>San Diego State University

- 9:20 Weighted Euclidean Balancing for a Matrix Exposure in Estimating Causal Effect  
*Juan Chen and ♦Yingchun Zhou.* East China Normal University
- 9:45 Inferring Causal Effects on Survival Probability in a Double-Conounded Setting  
<sup>♦</sup>*Yang Bai<sup>1</sup> and Yifan Cui<sup>2</sup>.* <sup>1</sup>National University of Singapore <sup>2</sup>Zhejiang University
- 10:10 Floor Discussion.

### Session 23CHI9: Some Recent Developments in Deep Learning

Room: CHANGNING HALL

Organizer: Jian Huang, The Hong Kong Polytechnic University.  
Chair: Jian Huang, The Hong Kong Polytechnic University.

- 8:30 Generative Conformal Prediction  
<sup>♦</sup>*Cheng Li<sup>1</sup>, Guohao Shen<sup>2</sup>, Yuanyuan Lin<sup>3</sup> and Jian Huang<sup>4</sup>.* <sup>1</sup>1155151973@link.cuhk.edu.hk  
<sup>2</sup>guohao.shen@polyu.edu.hk <sup>3</sup>ylin@cuhk.edu.hk  
<sup>4</sup>j.huang@polyu.edu.hk
- 8:55 Wasserstein Generative Regression  
*Tong Wang<sup>1</sup>, Shanshan Song<sup>1</sup>, Guohao Shen<sup>2</sup>, Yuanyuan Lin<sup>1</sup> and ♦Jian Huang<sup>2</sup>.* <sup>1</sup>The Chinese University of Hong Kong <sup>2</sup>The Hong Kong Polytechnic University
- 9:20 Nonparametric Estimation of Non-Crossing Quantile Regression Process with Deep Requ Neural Networks  
<sup>♦</sup>*Guohao Shen<sup>1</sup>, Yuling Jiao<sup>2</sup>, Yuanyuan Lin<sup>3</sup>, Joel L. Horowitz<sup>4</sup> and Jian Huang<sup>1</sup>.* <sup>1</sup>The Hong Kong Polytechnic University <sup>2</sup>Wuhan University <sup>3</sup>The Chinese University of Hong Kong <sup>4</sup>Northwestern University
- 9:45 Sparse Kronecker Product Decomposition: a General Framework of Signal Region Detection in Image Regression  
*Sanyou Wu and ♦Long Feng.* University of Hong Kong
- 10:10 Floor Discussion.

### July 3, 10:30-12:10

### Session 23CHI110: Recent Advances on Functional Data Analysis

Room: LONGFENG HALL-III

Organizer: Jin-Ting Zhang, National University of Singapore, Singapore.  
Chair: Tianming Zhu, Nanyang Technological University, Singapore.

- 10:30 New Anova Tests for Multivariate Functional Data with Applications  
<sup>♦</sup>*Zhiping Qiu, Jiangyuan Fan and Jin-Ting Zhang.*
- 10:55 Long-Memory log-Linear Zero-Inflated Generalized Poisson Autoregression for Covid-19 Pandemic Modeling  
<sup>♦</sup>*Xiaofei Xu<sup>1</sup>, Ying Chen<sup>2</sup>, Yan Liu<sup>3</sup>, Yuichi Goto<sup>4</sup> and Masanobu Taniguchi<sup>3</sup>.* <sup>1</sup>Wuhan University <sup>2</sup>National University of Singapore <sup>3</sup>Waseda University <sup>4</sup>Kyushu University

- 11:20 One-Way Manova for Functional Data via Lawley–hotelling Trace Test  
<sup>♦</sup>*Tianming Zhu<sup>1</sup>, Jin-Ting Zhang<sup>2</sup> and Ming-Yen Cheng<sup>3</sup>.*  
<sup>1</sup>National Institute of Education, Nanyang Technological University <sup>2</sup>National University of Singapore <sup>3</sup>Hong Kong Baptist University
- 11:45 A Unified Approach to Hypothesis Testing for Functional Linear Models  
<sup>♦</sup>*Yinan Lin and ♦Zhenhua Lin.* National University of Singapore
- 12:10 Floor Discussion.

### Session 23CHI112: Recent Advances in Variable Selection and Regression Analysis with Interval-Censored Data

Room: SONGJIANG HALL

Organizer: Peijie Wang, Jilin University.  
Chair: Peijie Wang, Jilin University.

- 10:30 Joint Modelling of Current Status Data and Competing Risks  
<sup>♦</sup>*Da Xu, Tao Hu and Jianguo Sun.*
- 10:55 Transformation Models with Latent Variables and Informative Partly Interval-Censored Data  
<sup>♦</sup>*Jingjing Jiang<sup>1</sup>, Chunjie Wang<sup>1</sup>, Pan Deng<sup>2</sup> and Xinyuan Song<sup>3</sup>.* <sup>1</sup>Changchun University of Technology <sup>2</sup>Huazhong University of Science and Technologygy <sup>3</sup>The Chinese University of Hong Kong
- 11:20 Variable Selection for Bivariate Interval-Censored Failure Time Data under Linear Transformation Models  
<sup>♦</sup>*Rong Liu<sup>1</sup>, Mingyue Du<sup>2</sup> and Jianguo Sun<sup>3</sup>.* <sup>1</sup>Chnaghcun University of Technology <sup>2</sup>Jilin University <sup>3</sup>University of Missouri
- 11:45 Simultaneous Variable Selection and Estimation for Interval-Censored Failure Time Data with Ancillary Information  
<sup>♦</sup>*Mingyue Du and Xingqiu Zhao.* The Hong Kong Polytechnic University
- 12:10 Floor Discussion.

### Session 23CHI124: Order-of-Addition Experimental Designs

Room: YANGPU HALL

Organizer: Yongdao Zhou, Nankai University.  
Chair: Shanqi Pang, Henan Normal University.

- 10:30 Design for Order-of-Addition Experiments with Two-Level Components  
*Hengzhen Huang.* Guangxi Normal University
- 10:55 Fast Approximation of the Shapley Values Based on Order-of-Addition Experimental Designs  
<sup>♦</sup>*Liuqing Yang<sup>1</sup>, Yongdao Zhou<sup>1</sup>, Haoda Fu<sup>2</sup>, Min-Qian Liu<sup>1</sup> and Wei Zheng<sup>3</sup>.* <sup>1</sup>Nankai University <sup>2</sup>Eli Lilly and Company <sup>3</sup>University of Tennessee
- 11:20 Blocked Designs with Multi Block Variables  
<sup>♦</sup>*Shengli Zhao<sup>1</sup>, Qianqian Zhao<sup>1</sup>, Yuna Zhao<sup>2</sup> and Minqian Liu<sup>3</sup>.* <sup>1</sup>Qufu Normal University <sup>2</sup>Shandong Normal University <sup>3</sup>Nankai University

- 11:45 Order-of-Addition Experiments on the Adjacency Relationship  
*Xinran Zhang, Ruonan Zheng, Min-Qian Liu and \*Jian-Feng Yang.* Nankai University
- 12:10 Floor Discussion.

### Session 23CHI146: Recent Developments in Complex Data Structure Analysis

Room: XUHUI HALL

Organizer: Yuehan Yang, Central University of Finance and Economics.

Chair: Yuehan Yang, Central University of Finance and Economics.

- 10:30 An Iterative Model-Free Feature Screening Procedure: Forward Recursive Selection  
*\*Siwei Xia<sup>1</sup> and Yuehan Yang<sup>2</sup>.* <sup>1</sup>School of Science, Civil Aviation Flight University of China <sup>2</sup>School of Statistics and Mathematics, Central University of Finance and Economics

- 10:55 Modeling Air Quality Level with Autoregression

*Fukang Zhu.* Jilin University

- 11:20 Dimension Reduction of High-Dimension Categorical Data with Two or Multiple Responses Considering Interactions Between Responses

*Yuehan Yang.* Central University of Finance and Economics

- 11:45 Capped Asymmetric Elastic Net Support Vector Machine for Robust Binary Classification

*\*Kai Qi and Hu Yang.* College of Mathematics and Statistics, Chongqing University, Chongqing, China

- 12:10 Floor Discussion.

### Session 23CHI150: Statistical Analysis of Massive Data and Network Data

Room: CHANGNING HALL

Organizer: Rui Pan, Central University of Finance and Economics.

Chair: Rui Pan, Central University of Finance and Economics.

- 10:30 Divide-and-Conquer Mcmc for Massive Stationary Time Series Data via Spectral Methods  
*Feng Li.*

- 10:55 A Degree-Corrected Cox Model for Dynamic Networks  
*Yuguo Chen, \*Lianqiang Qu, Jinfeng Xu, Ting Yan and Yunpeng Zhou.*

- 11:20 Community Detection of Discrete-Time Temporal Networks under Dynamic Stochastic Block Models  
*Binghui Liu.*

- 11:45 Grid Point Approximation for Distributed Nonparametric Smoothing and Prediction  
*\*Yuan Gao<sup>1</sup>, Rui Pan<sup>2</sup>, Feng Li<sup>2</sup>, Riquan Zhang<sup>1</sup> and Hansheng Wang<sup>3</sup>.* <sup>1</sup>East China Normal University <sup>2</sup>Central University of Finance and Economics <sup>3</sup>Peking University

- 12:10 Floor Discussion.

### Session 23CHI151: Statistical Learning for Large-Scale and Complex Data

Room: JINGAN HALL

Organizer: Lei Wang, Nankai University.

Chair: Lei Wang, Nankai University.

- 10:30 Robust Distributed Learning  
*Xiaozhou Wang.* East China Normal University
- 10:55 Regularized Spectral Clustering under the Mixed Membership Stochasticblock Model  
*Huan Qing<sup>1</sup> and \*Jingli Wang<sup>2</sup>.* <sup>1</sup>China University of Mining and Technology <sup>2</sup>Nankai University
- 11:20 Floor Discussion.

### Session 23CHI152: Causal Reinforcement Learning

Room: NANHUI HALL

Organizer: Jingyong Su, Harbin Institute of Technology.

Chair: Jingyong Su, Harbin Institute of Technology.

- 10:30 Dnet: Distributional Network for Distributional Individualized Treatment Effects  
*Shikai Luo.* ByteDance
- 10:55 Enhancing Spaced Repetition Scheduling Through Memory Dynamics Modelling and Stochastic Optimization  
*Junyao Ye.* MaiMemo Inc.
- 11:20 Simultaneous Feature Selection and Clustering Based on Square Root Optimization  
*\*He Jiang<sup>1</sup>, Shihua Luo<sup>2</sup> and Yao Dong<sup>2</sup>.* <sup>1</sup>Xi'an Jiao Tong University <sup>2</sup>Jiangxi University of Finance and Economics
- 11:45 Causal Inference and Recommendation System  
*Kaixian Yu.*
- 12:10 Floor Discussion.

### Session 23CHI172: Junior Researcher Award Session

Room: JIADING HALL

Organizer: Junior Researcher Award Committee.

Chair: Pengsheng Ji, University of Georgia.

- 10:30 Distributional Shift-Aware Off-Policy Interval Estimation  
*\*Wenzhuo Zhou<sup>1</sup>, Yuhang Li<sup>2</sup>, Ruqing Zhu<sup>2</sup> and Annie Qu<sup>1</sup>.*  
<sup>1</sup>University of California Irvine <sup>2</sup>University of Illinois Urbana Champaign
- 10:55 Probing Differential Expression Patterns Efficiently and Robustly Through Adaptive Linear Multi-Rank Two-Sample Tests  
*Dan Daniel Erdmann-Pham.* Stein Fellow, Statistics Department at Stanford University
- 11:20 Efficient Algorithms for Large-Scale Optimal Transport Problems  
*Cheng Meng.* Renmin University of China
- 11:45 Testing Serial Independence of Object-Valued Time Series  
*Feiyu Jiang.* Fudan university
- 12:10 Floor Discussion.

### Session 23CHI35: Non-Euclidean Data Analysis

Room: CHONGMING HALL

Organizer: Xueqin Wang, University of Science and Technology of China.

Chair: Xueqin Wang, University of Science and Technology of China.

- 10:30 Intrinsic Mave for Spd Matrices and Beyond  
*Baiyu Chen, Shuang Dai and ♦Zhou Yu.* East China Normal University
- 10:55 Testing Strict Stationarity of Complex Time Series  
 ♦*Qiang Zhang<sup>1</sup>, Wenliang Pan<sup>2</sup>, Jin Zhu<sup>3</sup> and Xueqin Wang<sup>4</sup>.* <sup>1</sup>Chengdu University <sup>2</sup>Chinese Academy of Sciences <sup>3</sup>Sun Yat-Sen University <sup>4</sup>University of Science and Technology of China
- 11:20 Fpls-Dc: Functional Partial Least Squares Through Distance Correlation for Imaging Genomics  
 ♦*Wenliang Pan<sup>1</sup>, Chuang Li<sup>2</sup>, Tengfei Li<sup>3</sup>, Yue Shan<sup>3</sup>, Yun Li<sup>3</sup> and Hongtu Zhu<sup>3</sup>.* <sup>1</sup>Academy of Mathematics and Systems Science, Chinese Academy of Sciences <sup>2</sup>Sun Yat-sen university <sup>3</sup>University of North Carolina at Chapel Hill
- 11:45 Ball Impurity: Measuring Heterogeneity in General Metric Spaces  
*Ting Li.* The Hong Kong Polytechnic University
- 12:10 Floor Discussion.

### Session 23CHI38: Recent Advances in Privacy-Protected Data Collection and Analysis

Room: LONGFENG HALL-I  
 Organizer: Samuel Wu, University of Florida.  
 Chair: Samuel Wu, University of Florida.

- 10:30 Bias Correction in Analysis of Matrix Masked Data  
 ♦*Linh Nghiem<sup>1</sup>, Adam Ding<sup>2</sup> and Samuel Wu<sup>3</sup>.* <sup>1</sup>University of Sydney <sup>2</sup>Northwestern University <sup>3</sup>University of Florida
- 10:55 Reducing Noise Level in Differentially Private Data Collection Through Matrix Masking  
 ♦*Aidong Adam Ding<sup>1</sup>, Samuel Wu, Guanhong Miao and Shigang Chen.* <sup>1</sup>Northeastern University (USA)
- 11:20 A General Differentially Private Learning Framework for Decentralized Data  
*Lingchen Kong.*
- 11:45 Floor Discussion.

### Session 23CHI42: Recent Developments in Clinical Trial Design and Data Analysis

Room: JINSHAN HALL  
 Organizer: Yisheng Li, University of Texas MD Anderson Cancer Center.  
 Chair: Yisheng Li, University of Texas MD Anderson Cancer Center.

- 10:30 Meta-Analytic Evaluation of Surrogate Endpoints in Randomized Controlled Trials with Varying Follow-Up Durations and Non-Proportional Hazards  
 ♦*Xiaoyu Tang<sup>1</sup> and Ludovic Trinquart<sup>2</sup>.* <sup>1</sup>Pfizer, Shanghai, China <sup>2</sup>Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, USA
- 10:55 Validation of Predictive Analyses for Interim Decisions in Clinical Trials  
*Lorenzo Trippa.* Dana-Farber Cancer Institute

- 11:20 A Semi-Mechanistic Dose-Finding Design in Oncology using Pharmacokinetic/Pharmacodynamic Modeling  
*Xiao Su<sup>1</sup>, ♦Yisheng Li<sup>2</sup>, Peter Müller<sup>3</sup>, Chia-Wei Hsu<sup>4</sup>, Haitao Pan<sup>4</sup> and Kim-Anh Do<sup>2</sup>.* <sup>1</sup>PlayStation <sup>2</sup>University of Texas MD Anderson Cancer Center <sup>3</sup>University of Texas at Austin <sup>4</sup>St. Jude Children's Research Hospital
- 11:45 Floor Discussion.

### Session 23CHI45: Recent Advances in Statistical Methodology

Room: NANSHI HALL  
 Organizer: Yuehua Wu, York University.  
 Chair: Yuehua Wu, York University.

- 10:30 Connecting Probability Estimation and Statistical Inference for Random Networks  
*Yichen Qin.* University of Cincinnati
- 10:55 Testing of Social Network Dependence Based on Autoregressive Model  
 ♦*Baisuo Jin, Wenyi Li and Xueqin Wang.*
- 11:20 Online Statistical Inference for Matrix Contextual Bandit  
*Qiyu Han, ♦Will Wei Sun and Yichen Zhang.* Purdue
- 11:45 On Robust Estimation of Hidden Semi-Markov Regime-Switching Model  
 ♦*Shanshan Qin<sup>1</sup>, Zhenyi Tan<sup>2</sup> and Yuehua Wu<sup>2</sup>.* <sup>1</sup>Tianjin University of Finance and Economics <sup>2</sup>York University
- 12:10 Floor Discussion.

### Session 23CHI5: Innovative Statistical Methods

Room: HONGKOU HALL  
 Organizer: Hao Zhang, Department of Biostatistics and Programming, China, Sanofi, Inc..  
 Chair: Hao Zhang, Department of Biostatistics and Programming, China, Sanofi, Inc..

- 10:30 Use of Bayesian Hierarchy Model in Pediatric and Rare Disease with Data Extrapolation  
*Hao Zhang and ♦Wei Xu.* Sanofi China
- 10:55 The Graphical Approach with a Bonferroni Mixture of Weighted Simes Tests: a Case Study  
 ♦*Wei Zhang and Kaifeng Lu.* Beigene
- 11:20 Dealing with Missing Data in Multi-Reader Multi-Case Design Studies  
 ♦*Zhemin Pan<sup>1</sup>, Yingyi Qin<sup>2</sup> and Jia He<sup>2</sup>.* <sup>1</sup>Tongji University <sup>2</sup>Second Military Medical School
- 11:45 Quantifying the Investigator's Decision Bias in Clinical Trial: An Approach Based on Causal Mediation Inference  
 ♦*Bo Chen<sup>1</sup>, Xing Zhao<sup>2</sup> and Juying Zhang<sup>2</sup>.* <sup>1</sup>Shenzhen Chipscreen Biosciences CO. LTD; Sichuan University <sup>2</sup>Sichuan University
- 12:10 Floor Discussion.

### Session 23CHI78: Recent Developments in Survival Data Analysis

Room: BAOSHAN HALL  
 Organizer: Xingqiu Zhao, The Hong Kong Polytechnic University.  
 Chair: Xingqiu Zhao, The Hong Kong Polytechnic University.

---

Scientific Program (<sup>♦</sup>*Presenting author when there are multiple authors*)

---

- 10:30 Posterior Sampling from the Spiked Models via Diffusion Processes  
♦*Yuchen Wu and Andrea Montanari*. Stanford University
- 10:55 Model-Free Conditional Screening for Ultrahigh-Dimensional Survival Data via Conditional Distance Correlation  
*Hengjian Cui<sup>1</sup>, Yanyan Liu<sup>2</sup>, Guangcai Mao<sup>3</sup> and ♦Jing Zhang<sup>4</sup>.* <sup>1</sup>Capital Normal University <sup>2</sup>Wuhan University <sup>3</sup>Central China Normal University <sup>4</sup>Zhongnan University of Economics and Law
- 11:20 Rank-Based Greedy Model Averaging for High-Dimensional Survival Data  
♦*Baihua He<sup>1</sup>, Shuangge Ma<sup>2</sup>, Xinyu Zhang<sup>3</sup> and Lixing Zhu<sup>4</sup>*. <sup>1</sup>International Institute of Finance, School of Management, University of Science and Technology of China <sup>2</sup>Department of Biostatistics, Yale University <sup>3</sup>Academy of Mathematics and Systems Science, Chinese Academy of Sciences <sup>4</sup>Center for Statistics and Data Science, Beijing Normal University at Zhuhai
- 11:45 Cox Proportional Hazards Cure Model for Incomplete Auxiliary Covariate Data  
*Guangcai Mao*. Central China Normal University
- 12:10 Floor Discussion.

**Session 23CHI82: Advances in Statistical Methods and Inference for Complex Biomedical Data**

- Room: LONGFENG HALL-II  
Organizer: Yuedong Wang, University of California - Santa Barbara.  
Chair: Yuedong Wang, University of California - Satna Barbara.
- 10:30 Individualized Dynamic Model for Multi-Resolution Data  
*Jiuchen Zhang<sup>1</sup>, Fei Xue<sup>2</sup>, Qi Xu<sup>1</sup>, Jung-Ah Lee<sup>1</sup> and ♦Annie Qu<sup>1</sup>.* <sup>1</sup>UCI <sup>2</sup>Purdue U
- 10:55 Statistical Inference for Regression Models with a Diverging Number of Covariates: Beyond Linear Regression  
*Lu Xia<sup>1</sup>, Bin Nan<sup>2</sup> and ♦Yi Li<sup>3</sup>.* <sup>1</sup>Univ of Washington <sup>2</sup>UCI <sup>3</sup>Univ of Michigan
- 11:20 Use of Electronic Health Records (Ehr) Data for Research: Challenges and Opportunities  
*Hulin Wu*. University of Texas Health Science Center at Houston
- 11:45 Estimating the Reciprocal of a Binomial Proportion  
♦*Jiajin Wei<sup>1</sup>, Ping He<sup>2</sup> and Tiejun Tong<sup>1</sup>.* <sup>1</sup>Department of Mathematics, Hong Kong Baptist University <sup>2</sup>Faculty of Science and Technology, BNU-HKBU United International College
- 12:10 Floor Discussion.

# Abstracts

---

## **Session 23CHIKT1: Keynote Lecture 1**

### **Methods for Controlled Variable Selection in Linear, Generalized-Linear, and Index Models**

*Jun Liu*

Harvard University

A classical statistical idea is to introduce data perturbations and examine their impacts on a statistical procedure. In the same token, the knockoff methods carefully create imatching fake variables in order to measure how real signals stand out. I will discuss some recent investigations we made regarding both methodology and theory on a few related methods applicable to a wide class of regression models including the knock off filter, data splitting (DS), Gaussian mirror (GM), for controlling false discovery rate (FDR) in fitting linear, generalized linear and index models. We theoretically compare, under the weak-and-rare signal framework for linear models, how these methods compare with the oracle OLS method. We then focus on the DS procedure and its variation, Multiple Data Splitting (MDS), which is useful for stabilizing the selection result and boosting the power. DS and MDS are straightforward conceptually, easy to implement algorithmically, and applicable to a wide class of linear and nonlinear models. Interestingly, their specializations in GLMs result in scale-free procedures that can circumvent difficulties caused by non-traditional asymptotic behaviors of MLEs in moderate-dimensions and debiased Lasso estimates in high-dimensions. For index models, we had developed an earlier LassoSIR algorithm (Lin, Zhao and Liu 2019), which fits the DS framework quite well. I will also discuss some applications and open questions. The presentation is based on joint work with Chenguang Dai, Buyu Lin, Xin Xing, Tracy Ke, Yucong Ma, and Zhigen Zhao.

## **Session 23CHI100: Challenges in the Analysis of Survival Data with Complex Features**

### **Separable Pathway Effects of Semi-Competing Risks via Multi-State Models**

*Yuhao Deng<sup>1</sup>, ♦Yi Wang<sup>2</sup>, Xiang Zhan<sup>2</sup> and Xiao-Hua Zhou<sup>3</sup>*

<sup>1</sup>School of Mathematical Sciences, Peking University

<sup>2</sup>Beijing International Center for Mathematical Research, Peking University

<sup>3</sup>Beijing International Center for Mathematical Research, and Department of Biostatistics, School of Public Health, Peking University

wangyi@bicmr.pku.edu.cn

Semi-competing risks refer to the phenomenon where a primary outcome (such as mortality) can truncate an intermediate event (such as relapse of a disease), but not vice versa. Under the multi-state model, the primary outcome is decomposed to a direct outcome event and an indirect outcome event through intermediate events. Within this framework, we show that the total treatment effect on the cumulative incidence of the primary event can be decomposed into three separable pathway effects, corresponding to treatment effects on population-level transition rates between

states. We next propose estimators for the counterfactual cumulative incidences of the primary outcome under hypothetical treatments by generalized Nelson-Aalen estimators with inverse probability weighting, and then derive asymptotic properties of these estimators. Finally, we propose hypothesis testing procedures on these separable pathway effects based on logrank statistics. We have conducted extensive simulation studies to demonstrate the validity and superior performance of our new method compared with existing methods. As an illustration of its potential usefulness, the proposed method is applied to compare the effects of different allogeneic stem cell transplantation types on overall survival after transplantation.

### **Zero-Inflated Poisson Models with Measurement Error in Response**

*Grace Yi*

University of Western Ontario

gyi5@uwo.ca

The analysis of zero-inflated count data is often based on a mixture model which facilitates excess zeros in combination with a Poisson distribution, and various inference methods have been proposed under such a model. Such analysis procedures, however, are challenged by the presence of measurement errors in count responses. We propose a new measurement error model to describe error-contaminated count data. We show that ignoring the measurement error effects in the analysis may generally lead to invalid inference results, and meanwhile, we identify situations where ignoring measurement error can still yield consistent estimators. Furthermore, we propose a Bayesian method to address the effects of measurement error under the zero-inflated Poisson model and discuss the identifiability issues. Numerical studies are conducted to evaluate the performance of the proposed method.

### **Analysis of the Cox Model with Longitudinal Covariates with Measurement Errors and Partly Interval Censored Failure Times, with Application to an Aids Clinical Trial**

*♦Yanqing Sun<sup>1</sup>, Qingning Zhou<sup>1</sup> and Peter Gilbert<sup>2</sup>*

<sup>1</sup>University of North Carolina at Charlotte, USA

<sup>2</sup>Fred Hutchinson Cancer Center and University of Washington  
yasun@unc.edu

Time-dependent covariates are often measured intermittently and with measurement errors. Motivated by the AIDS Clinical Trials Group (ACTG) 175 trial, this paper develops statistical inferences for the Cox model for partly interval censored failure times and longitudinal covariates with measurement errors. Assuming an additive measurement error model for a longitudinal covariate, we propose a nonparametric maximum likelihood estimation approach by deriving the measurement error induced hazard model that shows the attenuating effect of using the plug-in estimate for the true underlying longitudinal covariate. An EM algorithm is devised to facilitate maximum likelihood estimation that accounts for the partly interval censored failure times. The proposed methods can accommodate different numbers of replicates for different individuals and at different times. Simulation studies show that the proposed methods perform well with satisfactory finite-sample performances and that the naive methods ignoring measurement error or using the plug-in estimate can yield large biases. A hypothesis testing procedure for the measurement error model is proposed. The proposed meth-

## Abstracts

---

ods are applied to the ACTG 175 trial to assess the associations of treatment arm and time-dependent CD4 cell count on the composite clinical endpoint of AIDS or death.

### Multi-Task Prediction Model for Survival Data

*Shuai You<sup>1</sup>, Xiaowen Cao<sup>1</sup>, Grace Yi<sup>2</sup>, ♦Xuekui Zhang<sup>1</sup> and Li Xing<sup>3</sup>*

<sup>1</sup>University of Victoria

<sup>2</sup>Western University

<sup>3</sup>University of Saskatchewan  
ubcxzhang@gmail.com

We proposed a multi-task prediction method, MTPS. MTPS simultaneously predicts multiple outcomes, and allows these predictions to share information to improve their performances. The original version of MTPS allows for predicting mixed types of outcomes including continuous and binary. In this talk, we briefly introduce MTPS and our recent work in extending it to handle survival outcome variables.

### Session 23CHI115: Recent Advanced of Modeling to Complex Censored Data

#### Regression Analysis for the Semi-Competing Models of Interval-Censored Data with a Cured Subgroup

*Yichen Lou<sup>1</sup>, ♦Peijie Wang<sup>1</sup> and Jianguo Sun<sup>2</sup>*

<sup>1</sup>Jilin University

<sup>2</sup>University of Missouri  
wangpj19890114@163.com

Semi-competing risks data occur frequently in medical research when interest is in the simultaneous modeling of two or more processes, one of which may censor the others. It occurs in many areas such as health science research and many methods for their analyses have been proposed. In particular, some methods have been developed for the situation with the existence of a cured subgroup under the right censoring. However, health sciences research often involves interval-censored failure time data because the occurrence of progressive disease can only be detected through periodic examinations, which is a lack of research. In this paper, we discuss the case where both a cured subgroup and interval-censored semi-competing risks data exist and a parametric and a sieve mixture cure model approach is proposed. An extensive simulation study is conducted and indicates that the proposed procedure works well in practice. In addition, the methodology is applied to a real study of dementia and death.

#### Joint Analysis of Informatively Interval-Censored Failure Time and Panel Count Data

*♦Shuying Wang<sup>1</sup>, Chunjie Wang<sup>1</sup>, Xinyuan Song<sup>2</sup> and Da Xu<sup>3</sup>*

<sup>1</sup>Changchun University of Technology

<sup>2</sup>The Chinese University of Hong Kong

<sup>3</sup>Northeast Normal University  
wangshuying0601@163.com

Interval-censored failure time and panel count data, which frequently arise in medical studies and social sciences, are two types of important incomplete data. Although methods for their joint analysis have been available in the literature, they did not consider the observation process, which may depend on the failure time and/or panel count of interest. This study considers a three-component joint model to analyze interval-censored failure time, panel counts, and the observation process within a unique framework. Gamma and distribution-free frailties are introduced to jointly model the interdependency among the interval-censored data, panel count data,

and the observation process. We propose a sieve maximum likelihood approach coupled with Bernstein polynomial approximation to estimate the unknown parameters and baseline hazard function. The asymptotic properties of the resulting estimators are established. An extensive simulation study suggests that the proposed procedure works well for practical situations. An application of the method to a real-life dataset collected from a cardiac allograft vasculopathy study is presented.

#### Smoothed Estimation on Optimal Treatment Regime in Semi-Supervised Setting

*Xiaoqi Jiao, ♦Mengjiao Peng and Yong Zhou*

East China Normal University

A treatment regime refers to the process of assigning the most suitable treatment to a patient based on their observed information. However, existing research on treatment regimes mainly relies on labeled data, which may result in the loss of valuable information from unlabeled data like historical data. To address this issue, we propose a semi-supervised framework for estimating the optimal treatment regime in a model-free setting by leveraging unlabeled data. Our approach involves three essential steps. First, we utilize a single index model to perform dimension reduction, followed by kernel regression to impute the missing outcomes in the unlabeled data. Second, several types of semi-supervised value functions are proposed based on the imputed values and also combine the labeled part and unlabeled part. Third, the optimal treatment regime is derived by maximizing the semi-supervised value function. We establish the consistency and asymptotic normality of our proposed estimators. In addition, we introduce a perturbation resampling procedure to estimate the variance. Simulations confirm these favorable properties of introducing unlabeled data in the estimation of optimal treatment regimes.

#### Estimation and Variable Selection for Single-Index Models with Right Censored Data via Distance Covariance

*Xiaohui Yuan  
yuanxh@ccut.edu.cn*

TBC

### Session 23CHI118: Experimental Designs and their Applications

#### Revised Schematic Array

*Zuolu Hao and ♦Yu Tang*

Soochow University  
ytang@suda.edu.cn

Association scheme was first introduced by statisticians in connection with the design of experiments, and has been proved very useful in many fields, including permutation groups, graphs, and coding theory. An array is called schematic if its runs form an association scheme with respect to distance. Schematic arrays, especially schematic orthogonal arrays, have been ideal tools used in designing experiments and generating software test suites. However, the definition of the original schematic array is too demanding. It not only requires the relationship between any two distinct rows, but also overemphasizes the single-row property. This drawback dramatically limits the existence results on schematic arrays. In this paper, we modify the original conditions of the association scheme and propose the concept of the revised schematic array. We further elaborate on the rationality of the revised definition. Finally, we also

provide two examples of revised schematic arrays, including three-level mirror-symmetric orthogonal arrays and Latin hypercube designs.

#### **Group-Orthogonal Subsampling for Big Data Linear Mixed Models**

*Jiaqing Zhu<sup>1</sup>, Lin Wang<sup>2</sup> and ♦Fasheng Sun<sup>1</sup>*

<sup>1</sup> NENU

<sup>2</sup> PurdueU

sunfs359@nenu.edu.cn

Linear mixed model is a popular and common modeling method in statistical analysis. It is computationally difficult to obtain parameter estimates in linear mixed model for big data. The current subsampling methods are mainly aimed at the situation where the data is independent, without considering the correlation within the data. We provide some theoretical results on information matrix for linear mixed model. Based on these findings, an optimal subsampling method for linear mixed model is proposed, which maximizes the determinant of the variance-covariance matrix of the subsampling estimator. Besides, the proposed subsampling procedure is also optimal under A-optimality criterion, which minimizes the trace of the variance-covariance matrix of the subsampling estimator. Furthermore, asymptotic property of the subsampling estimator is established. Numerical examples based on both simulated and real data are provided to illustrate the proposed subsampling method.

#### **Sequentially Weighted Uniform Designs**

*Yao Xiao<sup>1</sup>, Shiqi Wang<sup>2</sup>, Hong Qin<sup>1</sup> and ♦Jianhui Ning<sup>2</sup>*

<sup>1</sup> Zhongnan University of Economics and Law

<sup>2</sup> Central China Normal University

jhnning@ccnu.edu.cn

Uniform designs seek to distribute design points uniformly in the experimental domain. Some discrepancies have been developed to measure the uniformity by treating all factors equally. It is reasonable when there exists no prior information about the system or when the potential model is completely unclear. However, in the situation of sequential designs, experimental information, such as the importance of each factor, would be obtained from previous stage experiments. With this fact, the weighted L2-discrepancy is more suitable than the original discrepancy for choosing follow-up designs. In this paper, the sequentially weighted uniform design is proposed, which is obtained by minimizing the weighted L2-discrepancy. The weights, indicating the relative importance of each factor, are estimated through a Bayesian hierarchical Gaussian process method based on serial experimental data. Results from several classic computer simulator examples, as well as a real application in circuit design, demonstrate that the performance of our new method surpasses that of its counterparts.

#### **Uniform Design with Prior Information of Factors under Weighted Wrap-Around L2-Discrepancy**

*Zujun Ou*

Jishou University

ozj9325@mail.ccnu.edu.cn

TBC

#### **Session 23CHI136: New Statistical Learning for High-Dimensional Data**

##### **Transfer Learning for High-Dimensional Quantile Regression via Convolution Smoothing**

*♦Yijiao Zhang and Zhongyi Zhu*

Fudan University

20210690169@fudan.edu.cn

This paper studies the high-dimensional quantile regression problem under the transfer learning framework, where possibly related source datasets are available to make improvements on the estimation or prediction based solely on the target data. In the oracle case with known transferable sources, a smoothed two-step transfer learning algorithm based on convolution smoothing is proposed and the  $1/2$  estimation error bounds of the corresponding estimator are also established. To avoid including non-informative sources, we propose to select the transferable sources adaptively and establish its selection consistency under regular conditions. Monte Carlo simulations as well as an empirical analysis of gene expression data demonstrate the effectiveness of the proposed procedure.

##### **Center-Augmented $\ell_2$ -Type Regularization for Subgroup Learning**

*Huazhen Lin*

Southwestern University of Finance and Economics

linhz@swufe.edu.cn

The existing methods for subgroup analysis can be roughly divided into two categories: finite mixture models (FMM) and regularization methods with an  $\ell_1$ -type penalty. In this paper, by introducing the group centers and  $\ell_2$ -type penalty in the loss function, we propose a novel center-augmented regularization (CAR) method; this method can be regarded as a unification of the regularization method and FMM and hence exhibits higher efficiency and robustness and simpler computations than the existing methods. In particular, its computational complexity is reduced from the  $O(n^2)$  of the conventional pairwise-penalty method to only  $O(nK)$ , where  $n$  is the sample size and  $K$  is the number of subgroups. The asymptotic normality of CAR is established, and the convergence of the algorithm is proven. CAR is applied to a dataset from a multicenter clinical trial, Buprenorphine in the Treatment of Opiate Dependence; a larger  $R^2$  is produced and three additional significant variables are identified compared to those of the existing methods.

##### **Transfer Learning for High-Dimensional Quantile Regression via Convolution Smoothing**

*♦Yijiao Zhang and Zhongyi Zhu*

Fudan University

zhuzy@fudan.edu.cn

This paper studies the high-dimensional quantile regression problem under the transfer learning framework, where possibly related source datasets are available to make improvements on the estimation or prediction based solely on the target data. In the oracle case with known transferable sources, a smoothed two-step transfer learning algorithm based on convolution smoothing is proposed and the  $1/2$  estimation error bounds of the corresponding estimator are also established. To avoid including non-informative sources, we propose to select the transferable sources adaptively and establish its selection consistency under regular conditions. Monte Carlo simulations as well as an empirical analysis of gene expression data demonstrate the effectiveness of the proposed procedure. Key words and phrases: High-dimensional data; Quantile regression; Regularization; Smoothing; Transfer learning.

##### **Learning Individualized Minimal Clinically Important Difference (Imcid) from High-Dimensional Data**

*Jiwei Zhao*

University of Wisconsin Madison

jiwei.zhao@wisc.edu

Statistical significance has been widely used to infer the treatment effect in assessing the efficacy of a treatment or intervention; how-

## Abstracts

---

ever, there has been a growing recognition that statistical significance has its own limitations. Clinical significance, on the contrary, is usually desirable in practice as it provides a better assessment of the clinically meaningful improvement. A critical concept in evaluating clinical significance is minimal clinically important difference (MCID), the smallest change in the outcome that an individual patient would identify as important. In this talk, I will present a statistical learning framework for estimating the individualized MCID (iMCID) from high-dimensional data. In particular, I will present a path-following iterative algorithm and some novel nonregular theoretical results. Additionally, simulation studies that reinforce our theoretical findings and an application to the study of chondral lesions in knee surgery to demonstrate the usefulness of the proposed approach will also be discussed.

### Deflated Heteropca: Overcoming the Curse of Ill-Conditioning in Heteroskedastic Pca

*Yuchen Zhou and ♦Yuxin Chen*

University of Pennsylvania  
yuxinc@wharton.upenn.edu

This talk is concerned with estimating the column subspace of a low-rank matrix  $X$  from contaminated data. How to obtain optimal statistical accuracy while accommodating the widest range of signal-to-noise ratios (SNRs) becomes particularly challenging in the presence of heteroskedastic noise and unbalanced dimensionality. While the state-of-the-art algorithm HeteroPCA emerges as a powerful solution for solving this problem, it suffers from the curse of ill-conditioning, namely, its performance degrades as the condition number grows. In order to overcome this critical issue without compromising the range of allowable SNRs, we propose a novel algorithm, called Deflate-HeteroPCA, that achieves near-optimal and condition-number-free theoretical guarantees in terms of both  $L_2$  and  $L_{2,\inf}$  statistical accuracy. The proposed algorithm divides the spectrum of  $X$  into well-conditioned and mutually well-separated subblocks, and applies HeteroPCA to conquer each subblock successively. Further, an application of our algorithm and theory to two canonical examples – the factor model and tensor PCA – leads to remarkable improvement for each application.

### Session 23CHI165: Recent Advances in Scalable Bayesian Inference

#### Learnable Topological Features for Phylogenetic Inference

*Cheng Zhang*

Peking University  
chengzhang@math.pku.edu.cn

Structural information of phylogenetic tree topologies plays an important role in phylogenetic inference. However, finding appropriate topological structures for specific phylogenetic inference tasks often requires significant design effort and domain expertise. In this talk, we propose a novel structural representation method for phylogenetic inference based on learnable topological features. By combining the raw node features that minimize the Dirichlet energy with modern graph representation learning techniques, our learnable topological features can provide efficient structural information of phylogenetic trees that automatically adapts to different downstream tasks without requiring domain expertise. We demonstrate the effectiveness and efficiency of our method on simulated data and real data phylogenetic inference problems.

#### Bayesian Fixed-Domain Asymptotics for Covariance Parame-

#### ters in Spatial Gaussian Process Regression Models

♦*Cheng Li<sup>1</sup>, Saifei Sun<sup>1</sup> and Yichen Zhu<sup>2</sup>*

<sup>1</sup>National University of Singapore

<sup>2</sup>Duke University  
stalic@nus.edu.sg

Gaussian process models typically contain finite dimensional parameters in the covariance function that need to be estimated from the data. We study the Bayesian fixed-domain asymptotics for the covariance parameters in spatial Gaussian process regression models with an isotropic Matern covariance function, which has many applications in spatial statistics. For the model without nugget, we show that when the dimension of the domain is less than or equal to three, the microergodic parameter and the range parameter are asymptotically independent in the posterior. While the posterior of the microergodic parameter is asymptotically close in total variation distance to a normal distribution with shrinking variance, the posterior distribution of the range parameter does not converge to any point mass distribution in general. For the model with nugget, we derive new evidence lower bound and consistent higher-order quadratic variation estimators, which lead to explicit posterior contraction rates for both the microergodic parameter and the nugget parameter. We further study the asymptotic efficiency and convergence rates of Bayesian kriging prediction. All the new theoretical results are verified in numerical experiments and real data analysis.

#### Catalytic Priors: using Synthetic Data to Specify Prior Distributions in Bayesian Analysis

♦*Dongming Huang<sup>1</sup>, Feicheng Wang<sup>2</sup>, Donald Rubin<sup>2</sup> and Samuel Kou<sup>2</sup>*

<sup>1</sup>National University of Singapore

<sup>2</sup>Harvard University  
statdongminghuang@gmail.com

Catalytic prior distributions provide a general, easy-to-use, and interpretable approach to specify prior distributions for Bayesian analysis. These distributions are especially useful when observed data are insufficient for accurately estimating a complex target model. In this context, a catalytic prior distribution stabilizes a high-dimensional working model by shrinking it towards a simplified model. The shrinkage is achieved by supplementing the observed data with a small amount of synthetic data, generated from a predictive distribution under the simpler model. Catalytic priors have simple interpretations and are easy to formulate. We apply this framework to generalized linear models and propose various strategies for specifying a tuning parameter that governs the degree of shrinkage. In our numerical experiments and a real-world study, the performance of the inference based on the catalytic prior is superior to or comparable to that of other commonly used prior distributions.

#### Sampling with Constraints using Variational Methods

*Xin Tong*

National University of Singapore  
xin.t.tong@nus.edu.sg

Sampling-based inference and learning techniques, especially Bayesian inference, provide an essential approach to handling uncertainty in machine learning (ML). As these techniques are increasingly used in daily life, it becomes essential to safeguard the ML systems with various trustworthy-related constraints, such as fairness, safety, interpretability. We propose a family of constrained sampling algorithms which generalize Langevin Dynamics (LD) and Stein Variational Gradient Descent (SVGD) to incorporate a moment constraint or a level set specified by a general nonlinear function. By exploiting the gradient flow structure of LD and SVGD, we derive algorithms for handling constraints, including a

primal-dual gradient approach and the constraint controlled gradient descent approach. We investigate the continuous-time mean-field limit of these algorithms and show that they have  $O(1/t)$  convergence under mild conditions.

## Session 23CHI17: Statistical Methods for Complex Longitudinal and Survival Data

### Efficient Algorithms for Survival Data with Multiple Outcomes using the Frailty Model

Xifen Huang, Jinfeng Xu and <sup>♦</sup>Yunpeng Zhou  
u3514104@connect.hku.hk

Survival data with multiple outcomes are frequently encountered in biomedical investigations. An illustrative example comes from Alzheimer's Disease Neuroimaging Initiative study where the cognitively normal subjects may clinically progress to mild cognitive impairment and/or Alzheimer's disease dementia. Transition time from normal cognition to mild cognitive impairment and that from mild cognitive impairment to Alzheimer's disease are expected to be correlated within subjects and the dependence is often accommodated by the frailty (random effects). Estimation in the frailty model unavoidably involves multiple integrations which may be intractable and hence leads to severe computational challenges, especially in the presence of high-dimensional covariates. In this paper, we propose efficient minorization–maximization algorithms in the frailty model for survival data with multiple outcomes. The alternating direction method of multipliers is further incorporated for simultaneous variable selection and homogeneity pursuit via regularization and fusion. Extensive simulation studies are conducted to assess the performance of the proposed algorithms. An application to the Alzheimer's Disease Neuroimaging Initiative data is also provided to illustrate their practical utilities.

### Semiparametric Regression Analysis of Interval-Censored Multi-State Data with an Absorbing State

<sup>♦</sup>Yu Gu, Donglin Zeng and Danyu Lin

University of North Carolina at Chapel Hill  
yugu@live.unc.edu

In studies of chronic diseases, the health status of a subject can often be characterized by a finite number of transient disease states and an absorbing state, such as death. The times of transitions among the transient states are ascertained through periodic examinations and thus interval-censored. The time of reaching the absorbing state is known or right-censored, with the transient state at the previous instant being unobserved. In this talk, I will present a general framework for analyzing such multi-state data through semiparametric proportional intensity models with random effects. Nonparametric maximum likelihood estimation and sieve estimation are combined for inference, and a stable EM algorithm is developed for computation based on Poisson data augmentation. The resulting estimators are shown to be consistent, asymptotically normal and efficient. I will demonstrate the performance of the proposed methods through simulation studies and provide an application to a cardiac allograft vasculopathy study.

### Sure Joint Screening for High Dimensional Cox's Proportional Hazards Model under the Case-Cohort Design

<sup>♦</sup>Yi Liu<sup>1</sup> and Gang Li<sup>2</sup>

<sup>1</sup>Ocean University of China

<sup>2</sup>University of California at Los Angeles  
liuyi@amss.ac.cn

This paper develops a sure joint feature screening method for the case-cohort design with ultrahigh dimensional covariates. Our method is based on a sparsity-restricted Cox's proportional hazards model. An iterative reweighted hard thresholding algorithm is proposed to approximate the sparsity-restricted pseudo-partial likelihood estimator for joint screening. We show rigorously that our method possesses the sure screening property, with the probability of retaining all relevant covariates tending to 1 as the sample size goes to infinity. Our simulation results demonstrate that the proposed procedure has substantially improved screening performance over some existing feature screening methods for the case-cohort design especially when some covariates are jointly correlated, but marginally uncorrelated, to the event time outcome. A real data illustration is provided on a breast cancer data with high dimensional genomic covariates. We have implemented the proposed method using Matlab and made it available to readers through Github.

### Federated Survival Analysis via Data Augmentation using Multi-Task Variational Autoencoder

Hong Wang  
Central South University  
wh@csu.edu.cn

Survival models finds vast applications in biomedical studies. However, survival data used to train these models are usually distributed, censored and facing a growing concern for data privacy. In addition to these issues, it is less commonly recognized that survival times are usually tailed. In this study, we attempt to tackle such challenges via a novel federated learning scheme. The proposed scheme aims to mitigate the censoring and tailed data problems via data augmentation using multi-task variational autoencoder(MVAE). Experimental results from extensive simulated and real world survival datasets have demonstrated the effectiveness of the proposed methodology.

## Session 23CHI2: Recent Developments in Network Analysis and High Dimensional Data

### Higher-Order Accurate Two-Sample Network Inference and Network Hashing

Meijia Shao<sup>1</sup>, Dong Xia<sup>2</sup>, <sup>♦</sup>Yuan Zhang<sup>1</sup>, Qiong Wu<sup>3</sup> and Shuo Chen<sup>4</sup>

<sup>1</sup>The Ohio State University

<sup>2</sup>Hong Kong University of Science and Technology

<sup>3</sup>University of Pennsylvania

<sup>4</sup>University of Maryland, Baltimore  
yzhanghf@stat.osu.edu

Two-sample hypothesis testing for comparing two networks is an important yet difficult problem. Major challenges include: potentially different sizes and sparsity levels; non-repeated observations of adjacency matrices; computational scalability; and theoretical investigations, especially on finite-sample accuracy and minimax optimality. In this article, we propose the first provably higher-order accurate two-sample inference method by comparing network moments. Our method extends the classical two-sample t-test to the network setting. We make weak modeling assumptions and can effectively handle networks of different sizes and sparsity levels. We establish strong finite-sample theoretical guarantees, including rate-optimality properties. Our method is easy to implement and computes fast. We also devise a novel nonparametric framework of offline hashing and fast querying particularly effective for maintaining and querying very large network databases. We demonstrate the

## Abstracts

---

effectiveness of our method by comprehensive simulations. We apply our method to two real-world data sets and discover interesting novel structures.

### Ranking Inferences Based on the Top Choice of Multiway Comparisons

Jianqing Fan<sup>1</sup>, Zhipeng Lou<sup>1</sup>, ♦Weichen Wang<sup>2</sup> and Mengxin Yu<sup>1</sup>

<sup>1</sup>Princeton University

<sup>2</sup>The University of Hong Kong  
weichenw@hku.hk

This paper considers ranking inference of n items based on the observed data on the top choice among M randomly selected items at each trial. This is a useful modification of the Plackett-Luce model for M-way ranking with only the top choice observed and is an extension of the celebrated Bradley-Terry-Luce model that corresponds to M=2. Under a uniform sampling scheme in which any M distinguished items are selected for comparisons with probability p and the selected M items are compared L times with multinomial outcomes, we establish the statistical rates of convergence for underlying n preference scores using both  $\ell_2$ -norm and  $\ell_\infty$ -norm, with the minimum sampling complexity. In addition, we establish the asymptotic normality of the maximum likelihood estimator that allows us to construct confidence intervals for the underlying scores. Furthermore, we propose a novel inference framework for ranking items through a sophisticated maximum pairwise difference statistic whose distribution is estimated via a valid Gaussian multiplier bootstrap. The estimated distribution is then used to construct simultaneous confidence intervals for the differences in the preference scores and the ranks of individual items. They also enable us to address various inference questions on the ranks of these items. Extensive simulation studies

### Spectral Analysis on Networks with Attributes

Wanjie Wang

National University of Singapore  
wanjie.wang@nus.edu.sg

Social network data record the connections between objects. In past decades, the study of social network data has been an important topic. Together with the observed connections, the profile of the object itself is usually also recorded. We call the data on the node-level profile covariates. Both the covariates and the network reflect the underlying data structure. In this talk, I will introduce our results that how the covariates will help to solve problems on the networks, such as community detection, and how the network helps with problems about the covariates, such as feature selection. We mainly focus on the spectral methods, which are computationally efficient in such problems. Hence, I will introduce our theoretical results on the control of the eigenvectors/eigenvalues when we combine both the networks and covariates. We believe such results are useful for further studies.

### Snr Estimation under High-Dimensional Linear Models

Xiaohan Hu and ♦Xiaodong Li

UC Davis  
xdgli@ucdavis.edu

Estimation of signal-to-noise ratios and residual variances in high-dimensional linear models has important applications including heritability estimation in bioinformatics. Random effects likelihood estimators have been widely used in practice for SNR estimation, and it is known to be consistent when the model is misspecified. In this talk, we aim to investigate the conditions on both the design matrix and the coefficient vector, such that asymptotic behaviors for this SNR estimator can be explicitly derived. We will stress tools

from random matrix theory and normal approximation of quadratic forms. For future work, extensions to method-of-moments, diverging aspect ratios, and linear models with feature groups will be briefly discussed. This is a joint work with my student Xiaohan Hu.

### Session 23CHI4: Hot and Special Biostatistical Considerations in HA Guidelines

#### Missing Data Handling for Time-to-Event Data under the Framework of Estimand

Bin Yao and ♦Lei Wang

Beigene  
slimewanglei@163.com

The framework of estimand has been increasingly adopted by sponsors after being recommended by ICH E9 (R1) addendum and agency guidelines. When target estimand uses a treatment policy to deal with intercurrent event (IE), all efforts should be made to collect data through and beyond IEs. Missing data may nevertheless occur due to subject's dropout from the study, which results in either non-informative or informative censoring. Non-informative censoring is commonly assumed in primary analysis. But missing-at-random assumption may not always be appropriate. Informative censoring could introduce substantial bias to the interpretation of results. Sensitivity analysis is commonly used to evaluate the robustness of missing data assumption. This presentation focuses on TPA via multiple imputation. Our goal is to impute the event time for patients with informative censoring in one arm or both arms. Several imputation models to generate time-to-event data is discussed. Analysis model is applied to each imputed data set, and the results is combined using Rubin's rules, producing the final estimate. Those models are compared under multiple simulation settings which mimic real world practice. The bias that arises from informative censoring is also investigated in simulation studies, given different informative censoring rates.

#### Covariate-Adjusted Analysis Targeting Marginal Estimand for Binary and Timetoevent Outcomes: Considerations and Examples

Xin Zhang

Pfizer Inc.  
xin.zhang6@pfizer.com

Covariate adjustment is a statistical analysis method with potential to improve precision and reduce the required sample size for randomized clinical trials. Examples of widely used methods for adjusted analysis are stratified analysis, logistic regression, and Cox PH regression. Those methods target conditional estimand, while for marginal estimand, the unadjusted analysis is still predominantly used, particularly for clinical trials with a binary or time-to-event outcome. There is an extensive literature on the statistical methods of covariate adjustment for marginal estimand. The recent FDA draft guidance on covariate adjustment also recommended several statistical methods, including standardization methods and semi-parametric methods. In this talk, we will evaluate the use of those methods and other potential methods (eg, targeted MLE methods) for the adjusted risk difference and hazard ratio for marginal estimand. Critical issues such as standard error estimation, variable selection and missing data/censoring will be discussed, and examples will be provided.

#### Concentration-Qtc Analysis to Support Ich E14 with a Real

**Case****Kong Xin**

Kong.Xin@Sanofi.com

Introduction to ICH E14 and Concentration-QTC model. The PK and ECG data obtained from a Phase I study characterized the effects of the drug and had demonstrated there were no effects on QT/QTc intervals and other ECG parameters. The results were adequate to waive a formal TQT trial for regulatory submission.

**Immuno-Bridging in Vaccine Development**♦*Iris Sun*<sup>1</sup> and *Lanfang Xia*<sup>2</sup><sup>1</sup>Department of Biostatistics and programming, China, Sanofi, Inc.<sup>2</sup>PPD, part of Thermo Fisher Scientific

iris.sun@sanofi.com

TBC

**Session 23CHI59: Recent Advances in Survey Statistics and Data Science****Loss Distribution of Arbitrary Credit Portfolios****Yimin Yang**

Loyal Trust Bank

yang.yimin@gmail.com

In this paper, the loss distribution for a credit portfolio with arbitrary Probability of Default (PD) distribution and non-constant asset correlations is derived under the single-factor framework. This can be viewed as the ultimate extension of the popular Vasicek single-factor loss model where the PD and the asset correlation are commonly assumed to be constants. This result also enables us to directly measure bank's credit risk concentration through its estimated capital which is a key regulatory requirement for a bank. We provide a direct connection between the Capital of an arbitrary credit portfolio and its asset correlation.

**Variable Selection on Survey Data with Missing Values***Yang Li*<sup>1</sup>, *Haoyu Yang*<sup>1</sup>, *Haochen Yu*<sup>1</sup>, *Hanwen Huang*<sup>2</sup> and ♦*Ye Shen*<sup>2</sup><sup>1</sup>Renmin University of China<sup>2</sup>University of Georgia

yeshen@uga.edu

Considering the inevitable correlation among different datasets within the same subject, we propose a framework of variable selection on multiply-imputed data with penalized weighted least squares (PWLS-MI). The methodological development is motivated by an epidemiological study of avian influenza A H7N9 (A/H7N9) patients, in which nearly half of the variables collected through epidemiological surveys are not fully observed. Multiple imputation is commonly adopted as a missing data processing method. However, it generates correlations among imputed values within the same subject across datasets. Recent work on variable selection for multiply-imputed data does not fully address such similarities. We propose PWLS - MI to incorporate the correlation when performing the variable selection. PWLS - MI can be considered as a framework for variable selection on multiply-imputed data as it allows various penalties. We use adaptive LASSO as an illustrative example. Extensive simulation studies are conducted to compare PWLS-MI with recently developed methods. The results suggest that the proposed approach generally outperforms existing approaches. PWLS-MI is shown to select variables with clinical relevance when applied to the A/H7N9 database.

**Robust Personalized Federated Learning with Sparse Penalization***Weidong Liu*<sup>1</sup>, *Xiaojun Mao*<sup>1</sup>, ♦*Xiaofei Zhang*<sup>2</sup> and *Xin Zhang*<sup>3</sup><sup>1</sup>Shanghai Jiao Tong University<sup>2</sup>Zhongnan University of Economics and Law<sup>3</sup>Iowa State University

zhangxf@zuel.edu.cn

Federated learning (FL) is an emerging topic thanks to its advantage in collaborative learning with distributed data. Due to the heterogeneity in the local data-generating mechanism, it is important to consider personalization when developing federated learning methods. In this work, we propose a personalized federated learning (PFL) method for addressing the robust regression problem. Specifically, we aim to learn the regression weight by solving a Huber loss with the sparse fused penalty. Additionally, we designed our personalized federated learning for robust and sparse regression (PerFL-RSR) algorithm to solve the estimation problem in the federated system efficiently. Theoretically, we show the convergence property of the proposed PerFL-RSR algorithm and then show that our proposed estimator is statistically consistent. Thorough experiments and real data analysis are conducted to corroborate the theoretical results of our proposed personalized federated learning method.

**Correlated Quantization for Distributed Mean Estimation — a Sampler's Perspective***Xiaojun Mao*<sup>1</sup>, ♦*Hengfang Wang*<sup>2</sup> and *Xiaofei Zhang*<sup>3</sup><sup>1</sup>Shanghai Jiao Tong University<sup>2</sup>Fujian Normal University<sup>3</sup>Zhongnan University of Economics and Law

hengfang@fjnu.edu.cn

Quantization has attracted much attention for handling the increasing demand for communication and data privacy protection with the growing data size. In this talk, we introduce a variance reduced correlated quantization scheme for data with bounded support for distributed mean estimation. We prove the reduction of the mean square error of the estimated mean from our proposed method for both fixed and randomized designs compared to the correlated quantization method, under different levels and dimensions scenarios. Several synthetic data experiments are conducted, which exhibit the merits of our proposed method. We further apply the proposed method to real-world data with different learning tasks and our method provides promising results.

**Session 23CHI62: Recent Developments in Change Point Analysis and Related Topics****Generalized Fiducial Inference for Gev Change-Point Model**♦*Xia Cai*, *Yaru Qiao* and *Shanshan Li*

Hebei University of Science and Technology

caixiatju@126.com

Generalized extreme value (GEV) distribution has received much attention and been widely used in medicine, engineering and meteorology. In this paper, identifying a shift in the parameters of GEV distribution of a single change point based on generalized fiducial inference (GFI) is considered. Instead of choosing an appropriate prior in Bayesian method, Jacobi function of the parameters is constructed. Then the density of the generalized fiducial distribution is derived using GFI approach. An adaptive Metropolis-Hastings within Gibbs algorithm is applied to obtain the estimation of the change-point location. Simulations are conducted to investigate the performance of the proposed approach based on GFI. Finally, the

## Abstracts

---

proposed method is applied to a real example to illustrate the detecting procedure.

### Integrative Learning of Linear Non-Gaussian Directed Acyclic Graphs

♦Xuanyu Li<sup>1</sup> and Qingzhao Zhang<sup>2</sup>

<sup>1</sup>University of Chinese Academy of Sciences

<sup>2</sup>Xiamen University

lixuanyu22@mails.ucas.ac.cn

We consider the problem of learning multiple related linear non-Gaussian directed acyclic graphs in high dimensional cases, where the graphs share the same sparse topological structure. To exploit this group sparsity condition, we use the Group Lasso method and distance covariance statistics to jointly reconstruct the topological layers of the directed acyclic graphs in a bottom-up fashion. We also establish the nonasymptotic consistency result of our approach under mild condition, and theoretically demonstrate the quantitative advantage of our integrative learning method over separate learning methods. The effectiveness of our proposed method is further supported by the numerical comparison against popular competitors in various simulated examples as well as a real dataset application.

### Confidence Intervals for Heterogeneity in Meta-Analysis of the Rare Binary Events Based on Empirical Likelihood-Type Methods

Sha Li<sup>1</sup>, ♦Weizhong Tian<sup>2</sup>, Xinmin Li<sup>1</sup> and Wei Ning<sup>3</sup>

<sup>1</sup>School of Mathematics and Statistics, Qingdao University

<sup>2</sup>College of Big Data and Internet, Shenzhen Technology University, Shenzhen

<sup>3</sup>Department of Mathematics and Statistics, Bowling Green State University  
tianweizhong@sztu.edu.cn

In meta-analysis, heterogeneity between independent studies is one of the important reference indicators for comprehensive analyses. It is usually described in terms of the variance between groups of the random-effect model. The existing methods are used to construct their confidence intervals based on the assumption of the normal distribution, but in reality, this assumption is often violated, especially in rare binary events. Recently, the method of the jackknife empirical likelihood was proposed to build confidence intervals, however, it tends to perform not good than parametric methods when the number of studies involved is small. Therefore, we propose to use the adjusted jackknife empirical likelihood, transformed jackknife empirical likelihood, and transformed adjusted jackknife empirical likelihood to construct confidence intervals for heterogeneity of the rare binary event. We investigate the performance of the proposed methods through the simulations based on different distributions, especially for skewed distributions and the small number of studies. Applications to real data with the process of the proposed methods is also provided.

### Monitoring Sequential Structural Changes in Penalized High-Dimensional Linear Models

Wei Ning

Bowling Green State University  
wning@bgsu.edu

In this talk, we introduce a novel procedure to monitor the structural changes in the penalized regression model for high-dimensional data sequentially. Our approach utilizes a given historical data set to perform both variable selection and estimation simultaneously. The asymptotic properties of the test statistics are established under the null and alternative hypotheses. The finite sample behavior of the monitoring procedure is investigated with simulation studies.

The proposed method is applied to a real data set to illustrate the detection procedure.

### Session 23CHI68: Mendelian Randomization: Causal Inference and Beyond

#### Adjusting for Genetic Confounders in Transcriptome-Wide Association Studies Leads to Reliable Detection of Causal Genes

Siming Zhao<sup>1</sup>, Wesley Crouse<sup>2</sup>, ♦Sheng Qian<sup>2</sup>, Kaixuan Luo<sup>2</sup>, Matthew Stephens<sup>3</sup> and Xin He<sup>2</sup>

<sup>1</sup>Department of Biomedical Data Science, Dartmouth Cancer Center, Dartmouth College

<sup>2</sup>Department of Human Genetics, University of Chicago

<sup>3</sup>Department of Human Genetics, Department of Statistics, University of Chicago  
shengqian@uchicago.edu

Many methods have been developed to leverage Expression Quantitative Trait Loci (eQTLs) to nominate candidate genes of complex traits, including colocalization analysis, transcriptome-wide association studies (TWAS), and Mendelian Randomization (MR)-based methods. All these methods, however, suffer from a key problem: when using the eQTLs of a gene to assess its role in a trait, nearby variants and nearby genetic components of expression of other genes can be correlated with the eQTLs of the test gene, while affecting the trait directly. These “genetic confounders” often lead to false discoveries. Our novel method, causal-TWAS (cTWAS), borrowed ideas from statistical fine-mapping, and allowed us to adjust all genetic confounders. In our simulations, we found that existing methods all suffered from high false positive rates. In contrast, cTWAS showed calibrated false positive rates while maintaining power. Application of cTWAS on several common traits discovered novel candidate genes. We have also extended the cTWAS model to accommodate multiple types of molecular QTL data. We found much higher power of identifying putative genes, when jointly analyzing eQTL, splicing QTL (sQTL) and methylation QTL (mQTL) data. In conclusion, cTWAS is a novel statistical framework to integrate molecular QTL and GWAS data, enabling reliable gene discoveries.

#### Likelihood-Based Mendelian Randomization Analysis with Automated Instrument Selection and Horizontal Pleiotropic Modeling

Xiang Zhou

University of Michigan  
xzhouph@umich.edu

Mendelian randomization (MR) is a common tool for identifying causal risk factors underlying diseases. Here, we present a method, MR with automated instrument determination (MRAID), for effective MR analysis. MRAID borrows ideas from fine-mapping analysis to model an initial set of candidate single-nucleotide polymorphisms that are in potentially high linkage disequilibrium with each other and automatically selects among them the suitable instruments for causal inference. MRAID also explicitly models both uncorrelated and correlated horizontal pleiotropic effects that are widespread for complex trait analysis. MRAID achieves both tasks through a joint likelihood framework and relies on a scalable sampling-based algorithm to compute calibrated P values. Comprehensive and realistic simulations show that MRAID can provide calibrated type I error control and reduce false positives while being more powerful than existing approaches. We illustrate the benefits of MRAID for an MR screening analysis across 645 trait pairs in

U.K. Biobank, identifying multiple lifestyle causal risk factors of cardiovascular disease-related traits.

#### **Identifying Gene by Environment Interactions, a Framework of Mr Approach**

Xiaofeng Zhu

Case Western Reserve University

xxz10@case.edu

There is a long-standing disagreement as to whether gene-environment interactions (GxE) contribute to phenotypic variations of complex traits. Part of the reason may be that very few interactions have been reported to date, presumably owing to the generally low statistical power to detect GxE by existing methods. Here, we present a new approach to detect GxE that shares substantial similarity to the Mendelian randomization framework, which has been widely applied to infer causal relationships between exposures and disease outcomes. Our approach improves statistical power over the direct test for GxE in some cases and is not sensitive to population structure and phenotype measurements. We will further illuminate the method by an application to search for the interactions with either smoking or alcohol consumption for serum lipids.

#### **Robust Multivariable Mendelian Randomization Based on Constrained Maximum Likelihood**

Zhaotong Lin, Haoran Xue and ♦Wei Pan

University of Minnesota

panxx014@umn.edu

Mendelian randomization (MR) is a powerful tool for causal inference using observational GWAS summary data. Compared to the more commonly used univariable MR (UVMR), multivariable MR (MVMR) not only is more robust to the notorious problem of genetic (horizontal) pleiotropy, but also estimates the direct effect of each exposure on the outcome after accounting for possible mediating effects of other exposures. Despite promising applications, there is a lack of studies on MVMR's theoretical properties and robustness in applications. In this work, we propose an efficient and robust MVMR method based on constrained maximum likelihood (cML), called MVMR-cML, with strong theoretical support. Extensive simulations demonstrate that MVMR-cML performs better than other existing MVMR methods while possessing the above two advantages over its univariable counterpart. An application to several large-scale GWAS summary datasets to infer causal relationships between 8 cardiometabolic risk factors and coronary artery disease (CAD) highlights the usefulness and some advantages of the proposed method. For example, after accounting for possible pleiotropic and mediating effects, triglyceride (TG), low-density lipoprotein cholesterol (LDL), and systolic blood pressure (SBP) had direct effects on CAD; in contrast, the effects of high-density lipoprotein cholesterol (HDL), diastolic blood pressure (DBP) and body height diminished.

#### **Session 23CHI69: Recent Advances in Adaptive Clinical Trial Design**

##### **Adaptive Promising Zone Two-Stage Design for a Trial with Binary Endpoint**

Guogen Shan

University of Florida

gshan@ufl.edu

Adaptive designs are increasingly used in clinical trials to assess the effectiveness of new drugs. For a single-arm study with binary

outcome, several adaptive designs were developed by using numerical search algorithms and the conditional power approach. The design based on numerical search algorithms is able to identify the global optimal design, but the computational intensity limits the usage of these designs. The conditional power approach searches for the optimal design without expensive computing time. In addition, promising zone strategy was proposed to move on drug development to the follow-up stages when the interim results are promising. We propose to develop two adaptive designs: one based on the conditional power approach, and the other based on the promising zone strategy. These two designs preserve type I and II error rates. It is preferable to satisfy the monotonic property for adaptive designs: the second stage sample size decreases as the first stage responses go up. We theoretically proof this important property for the two proposed designs. The proposed designs can be easily applied to real trials with limited computing resources.

#### **Session 23CHI76: Statistical Methods and Application**

##### **Additive Autoregressive Models for Matrix Valued Time Series**

Hong-Fan Zhang

Southwest Jiaotong University

zhanghongfan@swjtu.edu.cn

We develop additive autoregressive models (Add-ARM) for the time series data with matrix valued predictors. The proposed models assume separable row, column and lag effects of the matrix variables, attaining stronger interpretability when compared with existing bilinear matrix autoregressive models. We utilize the Gershgorin's circle theorem to impose some certain conditions on the parameter matrices, which make the underlying process strictly stationary. We also introduce the alternating least squares estimation method to solve the involved equality constrained optimization problems. Asymptotic distributions of the parameter estimators are derived. In addition, we employ hypothesis tests to run diagnostics on the parameter matrices. The performance of the proposed models and methods is further demonstrated through simulations and real data analysis.

##### **A Stable and Adaptive Polygenic Signal Detection Method Based on Repeated Sample Splitting**

♦Yanyan Zhao<sup>1</sup> and Lei Sun<sup>2</sup>

<sup>1</sup>Shandong University

<sup>2</sup>University of Toronto

yanyan.zhao@sdu.edu.cn

Focusing on polygenic signal detection in high dimensional genetic association studies of complex traits, we develop a stable and adaptive test for generalized linear models to accommodate different alternatives. To facilitate valid post-selection inference for high dimensional data, our study here adheres to the original sampling-splitting principle but does so, repeatedly, to increase stability of the inference. We show the asymptotic null distribution of the proposed test for both fixed and diverging number of variants. We also show the asymptotic properties of the proposed test under local alternatives, providing insights on why power gain attributed to variable selection and weighting can compensate for efficiency loss due to sample splitting. We support our analytical findings through extensive simulation studies and two applications. The proposed procedure is computationally efficient and has been implemented as the R package DoubleCauchy.

##### **A Bayesian Latent Subgroup Phase i/Ii Platform Design to Co-**

## Abstracts

---

### Develop Optimal Biological Doses for Multiple Indications

♦Rongji Mu<sup>1</sup>, Xiaojiang Zhan<sup>2</sup> and Ying Yuan<sup>3</sup>

<sup>1</sup>Shanghai Jiao Tong University

<sup>2</sup>Servier Pharmaceuticals

<sup>3</sup>The University of Texas MD Anderson Cancer Center

rjmu@sjtu.edu.cn

The platform trial refers to a new type of phase I/II trial which aims to determine the optimal biological doses simultaneously for different indications. Although patients with different indications who are enrolled in the trial treated by the same new agent, it is common for the targeted agent to be effective for patients with some types of cancer, but not others. We propose a Bayesian latent subgroup trial design to accommodate such treatment heterogeneity across indications. We assume that a cancer type may belong to the sensitive subgroup, which is responsive to the treatment, or the insensitive subgroup, which is not responsive to the treatment. Conditionally on the latent subgroup membership of the treatment arms, we model the toxicity and efficacy endpoint. At each interim, we update the toxicity and efficacy model, as well as the estimate of the utility, based on the observed data across indications to inform the indication-specific decision of dose escalation and de-escalation and identify the optimal biological dose for each indication. Simulation study shows that the proposed design has desirable operating characteristics, and that it provides an efficient approach to accelerate the development of combination therapies.

### Jackknife Empirical Likelihood for the Lower Mean Ratio

♦Lei Huang, Li Zhang and Yichuan Zhao

stahl@sjtu.edu.cn

Measuring economic inequality is a significant and meaningful topic in our social system. The Gini index and Pietra ratio are used by many people, but limited to reflecting the sampling distribution. In this paper, we study the interval estimates with another measure called the lower-mean ratio u. By using jackknife empirical likelihood (JEL), adjusted jackknife empirical likelihood (AJEL), mean jackknife empirical likelihood (MJEL), mean adjusted jackknife empirical likelihood (MAJEL), and adjusted mean jackknife empirical likelihood methods, we propose the interval estimator for u. We make a comparison for these methods in terms of the coverage probability and the average confidence interval length in the simulation study. The simulation results indicate that MAJEL performs the best among these methods for small sample sizes of the skewed distribution. For a small sample size of normal distribution, both JEL and MJEL show better performance than the other methods but MJEL is relatively time-consuming. The two real data set analyses illustrate the proposed methods.

### Session 23CHI81: Statistical Inferences for Complex Data

#### Inference for Arma Time Series with Mildly-Varying Trend

♦Yinghuai Yi, Zening Song and Lijian Yang

Tsinghua University

yyh21@mails.tsinghua.edu.cn

This work is mainly divided into two parts. In the first part, we propose a theoretically justified method to analyze time series consisting of an ARMA error term and a smooth trend function which could diverge to infinity. The trend is estimated by B-spline regression, and the maximum likelihood estimator based on residuals is shown to be oracally efficient in the sense that it is asymptotically as efficient as if the true trend function were known and then removed

to obtain the ARMA errors. In the second part, we give the concrete form of the asymptotically correct simultaneous confidence band for the trend function based on kernel estimation method. Simulation experiments corroborate the theoretical findings.

### Robust Regression using Probabilistically Linked Data

Raymond Chambers<sup>1</sup>, Enrico Fabrizi<sup>2</sup>, Maria Ranalli<sup>3</sup>, Nichola Salvati<sup>4</sup> and ♦Suojin Wang<sup>5</sup>

<sup>1</sup>University of Wollongong

<sup>2</sup>Università Cattolica del Sacro Cuore

<sup>3</sup>Università degli Studi di Perugia

<sup>4</sup>Università di Pisa

<sup>5</sup>Texas A&M University

sjwang@stat.tamu.edu

There is growing interest in a data integration approach to survey sampling, particularly where population registers are linked for sampling and subsequent analysis. The reason for doing this is simple: it is only by linking the same individuals in the different sources that it becomes possible to create a data set suitable for analysis. But data linkage is not error free. Many linkages are non-deterministic, based on how likely a linking decision corresponds to a correct match, i.e., it brings together the same individual in all sources. High quality linking will ensure that the probability of this happening is high. Analysis of the linked data should take account of this additional source of error when this is not the case. This is especially true for secondary analysis carried out without access to the linking information, i.e., the often confidential data that agencies use in their record matching. We describe an inferential framework that allows for linkage errors when sampling from linked registers. We focus on secondary analysis and linear regression modelling, including the important special case of estimation of subpopulation and small area means. In doing so we consider both robustness and efficiency of the resulting linked data inferences.

### Jackknife Empirical Likelihood for the Mean Difference of Two Zero-Inflated Skewed Populations

Faysal Satter<sup>1</sup> and ♦Yichuan Zhao<sup>2</sup>

<sup>1</sup>Lowe's Companies

<sup>2</sup>Georgia State University

yichuan@gsu.edu

In constructing a confidence interval for the mean difference of two independent populations, we may encounter the problem of having a low coverage probability when there are many zeros in the data, and the non-zero values are highly positively skewed. The violation of the normality assumption makes parametric methods inefficient in such cases. In this paper, jackknife empirical likelihood (JEL) and adjusted jackknife empirical likelihood (AJEL) methods are proposed to construct a nonparametric confidence interval for the mean difference of two independent zero-inflated skewed populations. The JEL and AJEL confidence intervals are compared with the confidence intervals by normal approximation and empirical likelihood proposed by Zhou and Zhou (2005). Simulation studies are performed to assess the new methods. Two real-life datasets are also used as an illustration of the proposed methodologies.

### Testing Linearity in Semi-Functional Partially Linear Regression Models

♦Yongzhen Feng<sup>1</sup>, Jie Li<sup>2</sup> and Xiaojun Song<sup>3</sup>

<sup>1</sup>Tsinghua University

<sup>2</sup>Renmin University of China

<sup>3</sup>Peking University

fyz20@mails.tsinghua.edu.cn

This paper proposes a Kolmogorov – Smirnov type statistic and a Cramer–von Mises type statistic to test linearity in semi-functional partially linear regression models. Our test statistics are based on a residual marked empirical process indexed by a randomly projected functional covariate, which is able to circumvent the “curse of dimensionality” brought by the functional covariate. The asymptotic properties of the proposed test statistics under the null, the fixed alternative, and a sequence of local alternatives converging to the null at the  $n^{1/2}$  rate are established. A straightforward wild bootstrap procedure is suggested to estimate the critical values that are required to carry out the tests in practical applications. Results from an extensive simulation study show that our tests perform reasonably well in finite samples. Finally, we apply our tests to the Tecator and AEMET datasets to check whether the assumption of linearity is supported by these datasets.

## Session 23CHI88: Intelligent Algorithms and Data Mining

### Session 23CHI129: Optimal Designs

#### Optimal Row-Column Designs

♦Zheng Zhou and Yongdao Zhou

Nankai University

zz1548070567@gmail.com

Row-column designs have been widely used in experiments involving double confounding. Among them, one that provides unconfounded estimation of all main effects and as many twofactor interactions as possible is preferred, and is called optimal. Most current work focusses on the construction of two-level row-column designs, while the corresponding optimality theory has been largely ignored. Moreover, most constructed designs contain at least one replicate of a full factorial design, which are not flexible as the number of factors increases. In this study, a theoretical framework is built up to evaluate the optimality of row-column designs with prime level. A method for constructing optimal row-column designs with prime level is proposed. Subsequently, optimal full factorial three-level row-column designs are constructed for any parameter combination. Optimal fractional factorial two-level and three-level row-column designs are also constructed for cost-saving.

#### Design Admissibility and De La Garza Phenomenon in Multi-Factor Experiments

Holger Dette, ♦Xin Liu<sup>1</sup> and Rong-Xian Yue<sup>2</sup>

<sup>1</sup>Donghua University

<sup>2</sup>Shanghai Normal University  
liuxin@dhu.edu.cn

The determination of an optimal design for a given regression problem is an intricate optimization problem, especially for models with multivariate predictors. Several authors have developed sufficient conditions for the existence of minimally supported designs in univariate models, where the number of support points of the optimal design equals the number of parameters. These results generalize the celebrated de la Garza phenomenon . We study a geometric characterization of the support points of an optimal design to provide sufficient conditions for the occurrence of the de la Garza phenomenon in models with multivariate predictors and characterize properties of admissible designs in terms of admissibility of designs in conditional univariate regression models.

#### Noncircular Designs for Controlling Border Effects under the

#### Interference Model: Gain with No Pain

♦Xiangshun Kong<sup>1</sup>, Jie Fu<sup>1</sup> and Wei Zheng<sup>2</sup>

<sup>1</sup>Beijing institute of technology

<sup>2</sup>University of Tennessee. Knoxville  
kongsunday@163.com

In many applications of block designs, the responses of plots could be affected by treatments in their neighboring plots. This leads to challenges in design construction since we need to consider the positions of the treatments within a block on top of the treatment composition as in traditional block designs. What makes it more complicated is the border effect on the two edge plots caused by potential environmental impacts outside the blocks. For the latter, [3] proposed adding two additional guarding plots next to the edge plots, for which we apply certain treatments to manually control the impacts. There have been extensive studies of designs under this setup, however, we observe that existing literature has been focusing on circular designs where the treatments applied to the border effects are the same as the edge plots on their opposite sides. We argue that such a structural restriction is not always mandated in most applications. We consider noncircular designs, where the two guarding plots are allowed to take any treatments by design. Optimal designs are constructed for both direct and total effects estimations. It is found that circular sequences do not appear in these new designs in most cases.

#### Equivalence Theorems for c and Da-Optimality for Linear Mixed Effects Models with Applications to Multi-Treatment Group Assignments in Healthcare

Xin Liu<sup>1</sup>, ♦Rong-Xian Yue<sup>2</sup> and Weng Kee Wong<sup>3</sup>

<sup>1</sup>College of Science, Donghua University, Shanghai 201620, China

<sup>2</sup>Department of Mathematics, Shanghai Normal University, Shanghai 200234, China

<sup>3</sup>Department of Biostatistics, University of California, Los Angeles, CA 90095-1772, USA  
yue2@shnu.edu.cn

We construct c and D\_A-optimal approximate designs for linear mixed models with group-specific treatment for estimating parameters or contrasts in the population parameters. We establish equivalence theorems to confirm optimality of these designs under a linear mixed model and provide illustrative application to find D, D\_A and c-optimal designs for polynomial and fractional polynomial models with multi-treatment group assignments. For more complex models, we briefly review metaheuristics and their potential applications to find various optimal designs, including optimal designs for problems considered here and their extensions.

## Session 23CHI131: Advances in Bioinformatics, Data Science, and Clinical Trial Design

#### A Novel Transcriptional Risk Score for Risk Prediction of Complex Human Diseases

♦Nayang Shan<sup>1</sup>, Yuhua Xie<sup>2</sup>, Shuang Song<sup>3</sup>, Wei Jiang<sup>2</sup>, Zuoheng Wang<sup>2</sup> and Lin Hou<sup>3</sup>

<sup>1</sup>Capital University of Economics and Business

<sup>2</sup>Yale School of Public Health

<sup>3</sup>Tsinghua University  
shanny2022@cueb.edu.cn

Recently polygenic risk score (PRS) has been successfully used in the risk prediction of complex human diseases. Many studies incorporated internal information, such as effect size distribution, or external information, such as linkage disequilibrium, functional annotation, and pleiotropy among multiple diseases, to optimize the

## Abstracts

---

performance of PRS. To leverage on multiomics datasets, we developed a novel flexible transcriptional risk score (TRS), in which messenger RNA expression levels were imputed and weighted for risk prediction. In simulation studies, we demonstrated that singletissue TRS has greater prediction power than LDpred, especially when there is a large effect of gene expression on the phenotype. Multitissue TRS improves prediction accuracy when there are multiple tissues with independent contributions to disease risk. We applied our method to complex traits, including Crohn's disease, type 2 diabetes, and so on. The singletissue TRS method outperformed LDpred and AnnoPred across the tested traits. The performance of multitissue TRS is traitdependent. Moreover, our method can easily incorporate information from epigenomic and proteomic data upon the availability of reference datasets.

### Mendelian Randomization for Causal Inference Accounting for Pleiotropy and Sample Structure using Genome-Wide Summary Statistics

♦Xianghong Hu<sup>1</sup>, Jia Zhao<sup>1</sup>, Zhixiang Lin<sup>2</sup>, Yang Wang<sup>1</sup>, Heng Peng<sup>3</sup>, Hongyu Zhao<sup>4</sup>, Xiang Wan<sup>5</sup> and Can Yang<sup>1</sup>

<sup>1</sup>The Hong Kong University of Science and Technology

<sup>2</sup>The Chinese University of Hong Kong

<sup>3</sup>Hong Kong Baptist University

<sup>4</sup>Yale School of Public Health

<sup>5</sup>Shen Zhen Research Institute of Big Data

maxhu@ust.hk

Mendelian Randomization (MR) is a valuable tool for inferring causal relationships among a wide range of traits using summary statistics from genome-wide association studies (GWASs). Existing summary-level MR methods often rely on strong assumptions, resulting in many false positive findings. To relax MR assumptions, ongoing research has been primarily focused on accounting for confounding due to pleiotropy. Here we show that sample structure is another major confounding factor, including population stratification, cryptic relatedness, and sample overlap. We propose a unified MR approach, MR-APSS, which (i) accounts for pleiotropy and sample structure simultaneously by leveraging genome-wide information; and (ii) allows to include more genetic variants with moderate effects as instrument variables (IVs) to improve statistical power without inflating type I errors. We first evaluated MR-APSS using comprehensive simulations and negative controls, and then applied MR-APSS to study the causal relationships among a collection of diverse complex traits. The results suggest that MR-APSS can better identify plausible causal relationships with high reliability. In particular, MR-APSS can perform well for highly polygenic traits, where the IV strengths tend to be relatively weak and existing summary-level MR methods for causal inference are vulnerable to confounding effects.

### Seamless Phase II/Iii Clinical Trials with Covariate Adaptive Randomization

Wei Ma<sup>1</sup>, Mengxi Wang<sup>2</sup> and ♦Hongjian Zhu<sup>3</sup>

<sup>1</sup>China Renmin University

<sup>2</sup>UTHealth

<sup>3</sup>AbbVie Inc.

hongjian.zhu@abbvie.com

There is an urgent need to evaluate new therapies in a time-sensitive and cost-effective manner. We propose the adaptive seamless phase II/Iii clinical trials with covariate adaptive randomization (CAR) to satisfy this need. CAR is one of the most popular designs in randomized controlled trials, enhancing covariance balance and ensuring valid treatment comparisons. However, it has several chal-

lenges: (1) the type I error rate of the commonly used Student's t-test following CAR can be inflated because of the seamless trials, but can also be decreased using CAR; (2) the complicated allocation mechanism induced by CAR causes extra difficulties to derive the asymptotic properties of a test procedure; and (3) previous theoretical studies of seamless trials rely mainly on the assumption of complete randomization, a procedure rarely used in real trials. We establish a theoretical foundation for adaptive seamless phase II/Iii trials with CAR. We also propose an approach that is easy to implement in order to control the type I error rate and improve the power when using Student's t-test. This important step will promote the application of this procedure.

### Session 23CHI134: Message Passing and Differential Privacy

#### On Reference Panel-Based Regularized Estimators in High-Dimensional Sparsity-Free Genetic Data Prediction

♦Buxin Su<sup>1</sup>, Qiang Sun<sup>2</sup>, Xiaochen Yang<sup>3</sup> and Bingxin Zhao<sup>1</sup>

<sup>1</sup>University of Pennsylvania

<sup>2</sup>University of Toronto

<sup>3</sup>Purdue University  
subuxin@sas.upenn.edu

Regularized estimators based on reference panels have been widely applied to the genetic prediction of complex traits, in part because they eliminate data privacy restrictions and reduce both computational and data transition costs. In these estimators, the covariance matrix of predictors is estimated from an external reference panel rather than from the original training data. We study reference panel-based L<sub>1</sub> and L<sub>2</sub> regularized estimators in a unified high-dimensional prediction framework without sparsity constraints. We uncover several key factors that determine the performance of reference panel-based estimators. Furthermore, we show that reference panel-based estimators are likely to be less accurate in high dimensions than traditional regularized estimators, which is primarily due to the smaller sample size of the reference panel. It is also evident from our study that the accuracy cost of using reference panel-based estimators will become higher as more training data is collected, which highlights the importance of building large-scale reference panels for genetic risk prediction. Our theoretical analysis is based on novel results in approximate message passing for estimating a general covariance matrix using a reference panel. We numerically evaluate our results based on extensive simulations and real data analysis in the UK Biobank database.

#### Identification and Estimation of Causal Inference with Confounders Missing not at Random

♦Jian Sun and Bo Fu

复旦大学

jsun19@fudan.edu.cn

Making causal inferences from observational studies can be challenging when confounders are missing not at random. In such cases, identifying causal effects is often not guaranteed. Motivated by a real example, we consider a treatment-independent missingness assumption under which we establish the identification of causal effects when confounders are missing not at random. We propose a weighted estimating equation (WEE) approach for estimating model parameters and introduce three estimators for the average causal effect, based on regression, propensity score weighting, and doubly robust estimation. We evaluate the performance of these

estimators through simulations, and provide a real data analysis to illustrate our proposed method.

#### Differential Private Data Release for Mixed-Type Data via Latent Factor Models

♦*Yanqing Zhang<sup>1</sup>, Qi Xu<sup>2</sup>, Niansheng Tang<sup>1</sup> and Annie Qu<sup>2</sup>*

<sup>1</sup>Yunnan University

<sup>2</sup>University of California Irvine  
zyqznl2010@126.com

Differential privacy is a particular data privacy preserving technology which can publish synthetic data or statistical analysis with a minimum disclosure of private information of individual record. The tradeoff between privacy-preserving and utility guarantee is always a challenge for differential privacy technology, especially for synthetic data generation. In this paper, we propose a differential private synthetic data algorithm for mixed-type data with correlation based on latent factor models. The proposed method can add a relatively small amount of noise to synthetic data under the same level of privacy protection while capturing correlation information. Moreover, the proposed algorithm can generate synthetic data preserving the same data type as mixed-type original data, which greatly improves the utility of synthetic data. The key idea of our method is to partially perturb the factor matrix to construct a synthetic data generation model, and to utilize link functions to ensure consistency of synthetic data type with original data. The proposed method can generate privacy-preserving synthetic data at low computation cost even when the original data is high-dimensional. In theory, we establish differentially private properties of the proposed method. Our numerical studies also demonstrate superb performance of the proposed method.

#### The Distribution of Lasso and Its Applications: Arbitrary Covariance

*Yuting Wei*

University of Pennsylvania  
ytwei@wharton.upenn.edu

The Lasso estimator is a commonly used regression method for high-dimensional regression models in which the number of covariates  $p$  is larger than the number of observations  $n$ . It is known that in the regime where the ratio  $n/p$  is a constant, the Lasso estimator has a non-trivial distribution that involves extra noise due to the under-sampling effect. In this work, we first characterize the exact distribution of the Lasso estimator for a general class of design matrices with arbitrary covariance structure. This exact characterization enables us to develop some interesting consequences in risk estimation, hypothesis testing, and model selection.

#### Session 23CHI144: Applications of Bayesian Methods in Educational Statistics

##### Bayesian Model Assessment in Joint Modeling of Response and Response Time&data under the Generalized Semi-Parametric Model

♦*Fang Liu<sup>1</sup> and Ming-Hui Chen<sup>2</sup>*

<sup>1</sup>Soochow University

<sup>2</sup>University of Connecticut  
fangliu@suda.edu.cn

Recently, several novel Bayesian model assessment criteria (Liu, Wang, et al., 2022), which are used to separately access the contribution of different sources of the data under joint model, are proposed under psychological background. Those methods are developed based on simple joint parametric model, for example, the joint

model of logistic and log-normal distributions, which leads to analytical derivations and computations. However, for more complicated model (e.g., semi-parametric model), it is still unknown on computing out those key quantities, which may refer to more complicated derivations in theory and calculation in software. In this paper, we propose more efficient skills and methodologies to calculate those assessment criteria based on the decomposition of deviance-information criterion (DIC) and the logarithm of the pseudo-marginal likelihood (LPML) under GORH model. Then, the effective Code are developed to assess the performance for those model assessment criteria. Also, those promising tools can be extend to other more complicated models easily using the same technique. Further, we conducted several simulation studies to examine the empirical performance and illustrated the application with the Programme for International Student Assessment (PISA) science data.

##### Bayesian Inference for Multidimensional Irt Models with Flexible Distributions

♦*Xue Zhang<sup>1</sup>, Chun Wang<sup>2</sup> and David Weiss<sup>3</sup>*

<sup>1</sup>Northeast Normal University

<sup>2</sup>University of Washington

<sup>3</sup>University of Minnesota  
zhangx815@nenu.edu.cn

Under item response theory (IRT) framework, misspecified latent distribution may lead to biased parameter estimations, such as normality assumption for a non-normal distribution. For unidimensional IRT models with non-normal latent trait distributions, MCMC methods mostly outperform the MML method, especially when sample size is small. The purpose of this study was to extend the MCMC algorithm with Davidian curve (MCMC-DC) to handle the multidimensional IRT models with flexible latent trait distributions (i.e., normal, skewed, and bimodal), and to explore the adaptive selection method of the best fit DC order during estimating parameters. The performance of the proposed method was illustrated via simulation studies and a real data example. Preliminary results indicated that the MCMC-DC method could fit normal and bimodal distributions well and skewed distributions reasonably well, and the method provided good estimates of item parameters.

##### A Sequential Bayesian Changepoint Detection Procedure for Aberrant Behaviors in Computerized Testing

♦*Jing Lu<sup>1</sup>, Chun Wang<sup>2</sup>, Jiwei Zhang<sup>1</sup> and Xue Wang<sup>1</sup>*

<sup>1</sup>Northeast Normal University

<sup>2</sup>University of Washington  
zhangjw713@nenu.edu.cn

Changepoints are abrupt variations in a sequence of data in statistical inference. In educational and psychological assessments, it is pivotal to properly differentiate examinees' aberrant behaviors from solution behavior to ensure test reliability and validity. In this paper, we propose a sequential Bayesian changepoint detection algorithm to monitor the locations of changepoints for response times in real time and, subsequently, further identify the types of aberrant behaviors in conjunction with response patterns. Two simulation studies were conducted to investigate the efficiency and accuracy of the proposed detection procedure in terms of identifying one or multiple change points at different locations. In addition to manipulating the number and locations of changepoints, two types of aberrant behaviors were also considered: rapid guessing behavior and cheating behavior. Simulation results indicate that ability estimates could be improved after removing responses from aberrant behaviors identified by our approach. Two empirical examples were analyzed to

## Abstracts

---

illustrate the application of proposed sequential Bayesian changepoint detection procedure.

### A Sequential Bayesian Changepoint Detection Procedure for Aberrant Behaviors in Computerized Testing

♦Jing Lu<sup>1</sup>, Chun Wang<sup>2</sup>, Jiwei Zhang<sup>1</sup> and Xue Wang<sup>1</sup>

<sup>1</sup>Northeast Normal University

<sup>2</sup>University of Washington

luj282@nenu.edu.cn

Changepoints are abrupt variations in a sequence of data in statistical inference. In educational and psychological assessments, it is pivotal to properly differentiate examinees' aberrant behaviors from solution behavior to ensure test reliability and validity. In this paper, we propose a sequential Bayesian changepoint detection algorithm to monitor the locations of changepoints for response times in real time and, subsequently, further identify the types of aberrant behaviors in conjunction with response patterns. Two simulation studies were conducted to investigate the efficiency and accuracy of the proposed detection procedure in terms of identifying one or multiple change points at different locations. In addition to manipulating the number and locations of changepoints, two types of aberrant behaviors were also considered: rapid guessing behavior and cheating behavior. Simulation results indicate that ability estimates could be improved after removing responses from aberrant behaviors identified by our approach. Two empirical examples were analyzed to illustrate the application of proposed sequential Bayesian changepoint detection procedure.

### Variational Bayesian Algorithm in Dina Model

Juntao Wan<sup>t</sup>

wangjt566@nenu.edu.cn

TBC

## Session 23CHI155: Statistical Frontiers in Spatial, Single-Cell, and Single-Molecule Multi-Omics Data

### Distributed Semi-Supervised Sparse Statistical Inference

Jiyuan Tu<sup>1</sup>, Weidong Liu<sup>2</sup>, ♦Xiaojun Mao<sup>2</sup> and Mingyue Xu<sup>3</sup>

<sup>1</sup>Shanghai University of Finance and Economics

<sup>2</sup>Shanghai Jiao Tong University

<sup>3</sup>Columbia University

maoxj@sjtu.edu.cn

This paper is devoted to studying the semi-supervised sparse statistical inference in a distributed setup. An efficient multi-round distributed debiased estimator, which integrates both labeled and unlabeled data, is developed. We will show that the additional unlabeled data helps to improve the statistical rate of each round of iteration. Our approach offers tailored debiasing methods for  $M$ -estimation and generalized linear model according to the specific form of the loss function. Our method also applies to a non-smooth loss like absolute deviation loss. Furthermore, our algorithm is computationally efficient since it requires only one estimation of a high-dimensional inverse covariance matrix. We demonstrate the effectiveness of our method by presenting simulation studies and real data applications that highlight the benefits of incorporating unlabeled data.

### Functional Calibration under Non-Probability Survey Sampling

♦Zhonglei Wang<sup>1</sup>, Xiaojun Mao<sup>2</sup> and Jae Kwang Kim<sup>3</sup>

<sup>1</sup>Xiamen University

<sup>2</sup>Shanghai Jiao Tong University

<sup>3</sup>Iowa State University

wangzl@xmu.edu.cn

Non-probability sampling is prevailing in survey sampling, but ignoring its selection bias leads to erroneous inferences. We offer a unified nonparametric calibration method to estimate the sampling weights for a non-probability sample by calibrating functions of auxiliary variables in a reproducing kernel Hilbert space. The consistency and the limiting distribution of the proposed estimator are established, and the corresponding variance estimator is also investigated. Compared with existing works, the proposed method is more robust since no parametric assumption is made for the selection mechanism of the non-probability sample. Numerical results demonstrate that the proposed method outperforms its competitors, especially when the model is misspecified. The proposed method is applied to analyze the average total cholesterol of Korean citizens based on a non-probability sample from the National Health Insurance Sharing Service and a reference probability sample from the Korea National Health and Nutrition Examination Survey.

### Semiparametric Regression Based on Quadratic Inference Function for Multivariate Failure Time Data with Auxiliary Information

Feifei Yan, Lin Zhu, ♦Yanyan Liu, Jianwen Cai and Haibo Zhou  
liuyy@whu.edu.cn

This paper deals with statistical inference procedure of multivariate failure time data when the primary covariate can be measured only on a subset of the full cohort but the auxiliary information is available. To improve efficiency of statistical inference, we use quadratic inference function approach to incorporate the intra-cluster correlation and use kernel smoothing technique to further utilize the auxiliary information. The proposed method is shown to be more efficient than those ignoring the intra-cluster correlation and auxiliary information and is easy to implement. In addition, we develop a chi-squared test for hypothesis testing of hazard ratio parameters. We evaluate the finite-sample performance of the proposed procedure via extensive simulation studies. The proposed approach is illustrated by analysis of a real data set from the study of left ventricular dysfunction.

## Session 23CHI166: Recent Developments in Recurrent Event Data and Panel Count Data Analysis

### Estimation and Inference for Fixed Center Effects on Panel Count Data

♦Weiwei Wang<sup>1</sup>, Yijun Wang<sup>1</sup> and Xiaobing Zhao<sup>2</sup>

<sup>1</sup>Zhejiang Gongshang University

<sup>2</sup>Zhejiang University of Finance and Economics

754177502@qq.com

In familial or multi-center studies, comparisons of outcomes across different centers are of interest. An extensive body of statistical models has been developed for various outcomes. However, most existing models apply only to simple data types. In this article, a fixed center effect proportional mean model is suggested to quantify center effects with respect to panel count data. When the number of centers is large, the traditional estimation methods that treat these center effects as categorical variables have many parameters to be estimated and thus may not be feasible to implement. In order to avoid including so many unknown variables, a new estimation procedure is proposed, where the center effects can be easily estimated by the center-specific ratio of observed to expected cumulative numbers of panel count data. The dimension of the estimated parameter space of the proposed procedure is only dependent on

the number of covariates, and it is computationally more efficient. Given some regularity conditions, the asymptotic properties of the proposed estimators are established. Extensive simulation studies are conducted to assess the finite-sample properties of the proposed estimators. Finally, the proposed method is applied to a real dataset from the China Health and Nutrition Study.

### **Regression Analysis of Mixed Panel Count Data with Dependent Observation Processes**

♦Lei Ge<sup>1</sup>, Jaihee Choi<sup>2</sup>, Hui Zhao<sup>3</sup>, Yang Li<sup>4</sup> and Jianguo Sun<sup>5</sup>

<sup>1</sup>Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, USA;

<sup>2</sup>Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, USA

<sup>3</sup>School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, People's Republic of China

<sup>4</sup>Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, USA

<sup>5</sup>Department of Statistics, University of Missouri, Columbia, MO, USA  
ge1@iu.edu

Event history data commonly occur in many areas and a great deal of literature on their analysis has been established. However, most of the existing methods apply only to a single type of event history data. Recently, several authors have discussed the analysis of mixed types of event history data and the existence of dependent observation processes is another issue that one often has to deal with in the analysis of event history data. This paper discusses regression analysis of mixed panel count data with dependent observation processes, which has not been addressed in the literature, and for the problem, an approximate likelihood estimation approach is proposed. For the implementation, an EM algorithm is developed and the proposed estimators are shown to be consistent and asymptotically normal. An extensive simulation study is performed to assess the performance of the proposed approach and indicates that it works well in practical situations. An application to a set of real data is provided.

### **An Ipw Additive Model for Panel Count Data with Dependent Observation Process and Terminal Event**

Ni Li

Hainan Normal University  
lini@hainnu.edu.cn

Panel count data are frequently encountered in follow-up studies such as clinical trials and sociological studies. Models about this type data usually assume the underlying recurrent event process is independent with the observation process. However, the independence assumption cannot be always guaranteed in practical applications, especially when they both are relevant with the covariates. In this paper, we develop a semiparametric additive model weighted by propensity score for analyzing panel count data in the presence of some terminal events. By introducing propensity score into the model, we try to reduce the confounding bias of parameter estimate caused by the dependent observation process. In particular, an IPW technique is used, and the asymptotic properties of the proposed estimators are established. Numerical simulations were conducted to evaluate the proposed procedures. Finally, we apply the methodologies to analyzing a set of skin cancer data.

### **A Double Exponential Gamma-Frailty Model for Clustered Survival Data**

♦Mengqi Xie<sup>1</sup>, Jie Zhou<sup>1</sup> and Lei Liu<sup>2</sup>

<sup>1</sup>School of Mathematics, Capital Normal University, Beijing 100048, China

<sup>2</sup>Division of Biostatistics, Washington University in St. Louis, MO 63110, USA  
2210502189@cnu.edu.cn

We propose a double exponential gamma-frailty model for clustered survival data. This model addresses the limitation of shared gamma-frailty models, where the marginal effects of covariates diminish over time. To estimate parameters, we utilize a sieve maximum likelihood approach and employ Bernstein polynomials for approximating nondecreasing cumulative baseline functions. The estimators' asymptotic properties are also provided. The proposed method is demonstrated through numerical simulations and survival data from a Diabetic Retinopathy Study (DRS).

### **Session 23CHI26: Current Topics in Biostatistics I**

#### **Bootstrap Averaging Reduces the Mean Square Error of Kinetic Maps Recovered from Pet Data**

♦Fengyun Gu<sup>1</sup>, Finbarr O'sullivan<sup>2</sup> and Qi Wu<sup>2</sup>

<sup>1</sup>North China Electric Power University

<sup>2</sup>University College Cork  
fengyungu@126.com

Positron emission tomography (PET) is an imaging technique that is prominently used in both clinical and research settings. Recent developments in the spatial and temporal resolution capabilities of PET scanners have substantially increased interest in mapping kinetic parameters from the large-scale datasets produced by dynamically acquired studies. Several schemes have been proposed for this problem. A recently developed image-domain bootstrapping technique has made assessment of the sampling distribution of imaged kinetic information practical. Its main use has been to map voxel-level uncertainty in kinetic variables. The work here examines its potential to improve the mean square error (MSE) performance of kinetic maps by averaging results from individual bootstrap samples.

#### **A Bayesian Non-Parametric Approach for Causal Mediation with a Post-Treatment Confounder**

Woojung Bae and ♦Michael Daniels

University of Florida  
daniels@ufl.edu

We propose a new Bayesian non-parametric (BNP) method for estimating the causal effects of mediation in the presence of a post-treatment confounder. We specify an enriched Dirichlet process mixture (EDPM) to model the joint distribution of the observed data (outcome, mediator, post-treatment confounder, treatment, and baseline confounders). For identifiability, we use the extended version of the standard sequential ignorability as introduced in Hong et al. (2022, *Biometrics*). The observed data model and causal identification assumptions enable us to estimate and identify the causal effects of mediation, i.e., the natural direct effects (NDE), and indirect effects (NIE). Our method enables easy computation of NDE and NIE for a subset of confounding variables and addresses missing data through data augmentation under the assumption of ignorable missingness. We conduct simulation studies to assess the performance of our proposed method. Furthermore, we apply this approach to evaluate the causal mediation effect in the Rural LITE trial, demonstrating its practical utility in real-world scenarios.

#### **Transparent Sequential Learning for Monitoring Sequential**

## Abstracts

---

### Processes

♦ Peihua Qiu and Xiulin Xie

University of Florida  
pqiu@ufl.edu

Machine learning methods have been widely used in process monitoring. For handling statistical process control (SPC) problems, conventional supervised machine learning methods would have some difficulties. For instance, a training dataset containing both in-control and out-of-control process observations is rarely available in SPC applications. In the SPC literature, there have been some existing discussions on how to handle the lack of out-of-control observations in the training data, using the one-class classification, artificial contrast, real-time contrast, and some other novel ideas. However, these approaches have their own limitations to handle SPC problems. In this paper, we extend the self-starting process monitoring idea that has been employed widely in modern SPC research to a general learning framework for monitoring processes with serially correlated data. Under the new framework, process characteristics to learn are well specified in advance, and process learning is sequential in the sense that the learned process characteristics keep being updated during process monitoring. The learned process characteristics are then incorporated into a control chart for detecting process distributional shift based on all available data by the current observation time. It is shown that this new method is more reliable and effective than some representative existing machine learning SPC approaches.

## Session 23CHI34: Optimization in Statistics

### Fast Consistent Best-Subset Selection in Generalized Linear Models

♦ Junxian Zhu<sup>1</sup>, Jin Zhu<sup>2</sup>, Borui Tang<sup>3</sup>, Xuanyu Chen<sup>2</sup> and Xueqin Wang<sup>3</sup>

<sup>1</sup>National University of Singapore

<sup>2</sup>Sun Yat-Sen University

<sup>3</sup>University of Science and Technology of China  
junxian@nus.edu.sg

Best-subset selection under generalized linear models (GLM) gives the most parsimonious model that sufficiently explains the variation of various responses, which is important and appealing for regression analysis. The obstacle of selecting the best subset in practice is that the computational efficiency and statistical guarantees are difficult to be achieved simultaneously. In this article, we aim to overcome this obstacle by developing a fast algorithm that can consistently select the best subset with high probability. Our algorithm includes two nested parts. First, to pick a preferable subset among all subsets of the same cardinality, a fast algorithm is designed and shown to have the screening property under regularity conditions. Next, we propose a generalized information criterion to select the cardinality in GLM and establish the consistency of model selection. Under mild conditions, the computational complexity of our algorithm is proportional to the sample size and dimensionality with high probability. Numerical studies on simulated and real data not only verify the computational and statistical properties of our method but also demonstrate its superiority on variable selection and coefficient estimation compared to the state-of-the-art methods.

### High Dimensional Portfolio Selection with Cardinality Constraints

♦ Yifeng Guo<sup>1</sup>, Jin-Hong Du<sup>2</sup> and Xueqin Wang<sup>3</sup>

<sup>1</sup>The University of Hongkong  
<sup>2</sup>Carnegie Mellon University

<sup>3</sup>University of Science and Technology of China  
gyf9712@gmail.com

The expanding number of assets offers more opportunities for investors but poses new challenges for modern portfolio management (PM). As a central plank of PM, portfolio selection by expected utility maximization (EUM) faces uncontrollable estimation and optimization errors in ultrahigh-dimensional scenarios. Past strategies for high-dimensional PM mainly concern only large-cap companies and select many stocks, making PM impractical. We propose a sample-average-approximation-based portfolio strategy to tackle the difficulties above with cardinality constraints. Our strategy bypasses the estimation of mean and covariance, the Chinese walls in high-dimensional scenarios. Empirical results on S & P 500 and Russell 2000 show that an appropriate number of carefully chosen assets leads to better out-of-sample mean-variance efficiency. On Russell 2000, our best portfolio profits as much as the equally-weighted portfolio but reduces the maximum drawdown and the average number of assets by 10% and 90%, respectively. The flexibility and the stability of incorporating factor signals for augmenting out-of-sample performances are also demonstrated. Our strategy balances the trade-off among the return, the risk, and the number of assets with cardinality constraints. Therefore, we provide a theoretically sound and computationally efficient strategy to make PM practical in the growing global financial market.

### Directed Community Detection with Network Embedding

♦ Jingnan Zhang<sup>1</sup>, Xin He<sup>2</sup> and Junhui Wang<sup>3</sup>

<sup>1</sup>University of Science and Technology of China

<sup>2</sup>Shanghai University of Finance and Economics, China

<sup>3</sup>The Chinese University of Hong Kong  
jnzhang@ustc.edu.cn

Community detection in network data aims at grouping similar nodes sharing certain characteristics together. Most existing methods focus on detecting communities in undirected networks, where similarity between nodes is measured by their node features and whether they are connected. In this article, we propose a novel method to conduct network embedding and community detection simultaneously in a directed network. The network embedding model introduces two sets of vectors to represent the out- and in-nodes separately, and thus allows the same nodes belong to different out-and in-communities. The community detection formulation equips the negative log-likelihood with a novel regularization term to encourage community structure among the nodes representations, and thus achieves better performance by jointly estimating the nodes embeddings and their community structures. To tackle the resultant optimization task, an efficient alternative updating scheme is developed. More importantly, the asymptotic properties of the proposed method are established in terms of both network embedding and community detection, which are also supported by numerical experiments on some simulated and real examples.

### Abess: a Fast Best-Subset Selection Library in Python and r

Jin Zhu

Department of Statistical Science, Sun Yat-Sen University  
zhuj37@mail2.sysu.edu.cn

We introduce a new library named abess that implements a unified framework of best-subset selection for solving diverse machine learning problems, e.g., linear regression, classification, and principal component analysis. Particularly, abess certifiably gets the optimal solution within polynomial time with high probability under the linear model. Our efficient implementation allows abess to attain the solution of best-subset selection problems as fast as or even 20x faster than existing competing variable (model) selection

toolboxes. Furthermore, it supports common variants like best subset of groups selection and ridge-regularized best-subset selection. The core of the library is programmed in C++. For ease of use, a Python library is designed for convenient integration with scikit-learn, and it can be installed from the Python Package Index (PyPI). In addition, a user-friendly R library is available at the Comprehensive R Archive Network (CRAN). The source code is available at: <https://github.com/abess-team/abess>.

## Session 23CHI44: New Developments in Large-Scale and High-Dimensional Inference

### Statistically Guided Divide-and-Conquer for Sparse Factorization of Large Matrix

*Kun Chen, Ruipeng Dong, Wanwan Xu and ♦Zemin Zheng  
zhengzm@ustc.edu.cn*

The sparse factorization of a large matrix is fundamental in modern statistical learning. The appeal of this factorization is owing to its power in discovering a highly-interpretable latent association network, either between samples and variables or between responses and predictors. However, many existing methods are either ad hoc without a general performance guarantee, or are computationally intensive, rendering them unsuitable for large-scale studies. We formulate the statistical problem as a sparse factor regression and tackle it with a divide-and-conquer approach. In the first stage of division, we consider both sequential and parallel approaches for simplifying the task into a set of co-sparse unit-rank estimation (CURE) problems, and establish the statistical underpinnings of these commonly-adopted and yet poorly understood deflation methods. In the second stage of division, we innovate a contended stagewise learning technique, consisting of a sequence of simple incremental updates, to efficiently trace out the whole solution paths of CURE. Our algorithm has a much lower computational complexity than alternating convex search, and the choice of the step size enables a flexible and principled tradeoff between statistical accuracy and computational efficiency.

### Asymptotics of the Spatial-Sign Based Estimators of Location and Scatter in High-Dimension

*Qinwen Wang  
FUDAN UNIVERSITY  
wqw@fudan.edu.cn*

In this talk, we will investigate limiting spectral properties of high-dimensional sample spatial-sign covariance matrix and Tyler's M estimator. The populations under study are general enough to include the popular independent components model and the family of elliptical distributions, with possibly known or unknown location vectors. Both the empirical spectral distributions and the central limit theorems for a class of linear spectral statistics of the two matrix ensembles are studied.

### Ranking and Selection in Large-Scale Inference of Heteroscedastic Units

*♦Bowen Gang<sup>1</sup>, Luella Fu<sup>2</sup>, Wenguang Sun<sup>3</sup> and Gareth James James<sup>4</sup>*

<sup>1</sup>Fudan University

<sup>2</sup>San Francisco State University

<sup>3</sup>Zhejiang University

<sup>4</sup>Emory University  
gangbowen02@163.com

Choosing candidates to whom a limited set of resources will be distributed is a pervasive dilemma. In multiple testing procedures that

can be used to choose such candidates, power is traditionally defined as the number or proportion of correctly selected non-null hypotheses. We propose a generalized power that allows researchers to better select more desirable testing units and propose a specific formulation to capture not only if a unit has been correctly categorized as null or alternative but also to better reward the detection of larger effect sizes. Our new empirical Bayes multiple testing framework rewards discovering not just significant but large effects while controlling type I error. Hence, the selection process is better able to incorporate effect size into selection. We provide theoretical guarantees for FDR control and power optimization as well as numeric evidence for the utility of a generalized power.

### Multi-Dimensional Domain Generalization with Low-Rank Structures

*♦Sai Li<sup>1</sup> and Linjun Zhang<sup>2</sup>*

<sup>1</sup>Renmin University of China

<sup>2</sup>Rutgers University  
saili@ruc.edu.cn

In health-related studies, certain sub-populations may be underrepresented, which presents a challenge for researchers seeking to understand the characteristics of these groups. In this talk, we tackle this challenge in linear models by organizing the regression vectors of all the sub-populations into a tensor. We formulate the domain generalization problem as a tensor completion task, allowing us to learn about sub-populations with limited or no available data. Unlike previous studies in tensor completion, our model accounts for complex missing patterns and correlation structures. Our proposed method is supported by theoretical guarantees and numerical studies demonstrating its efficiency.

## Session 23CHI49: New Development of Statistical Methods for Genomic, Epigenomic, and Microbiome Data

### Pan-Cancer Analysis of Pathway-Based Gene Expression Pattern at the Individual Level Reveals Biomarkers of Clinical Prognosis

*Zhaohui Qin  
Emory University  
zhaohui.qin@emory.edu*

Identifying biomarkers to predict the clinical outcomes of individual patients is a fundamental problem in clinical oncology. Multiple single-gene biomarkers have already been identified and used in the clinics. However, multiple oncogenes or tumor-suppressor genes are involved during the process of tumorigenesis. Additionally, the efficacy of single-gene biomarkers is limited by the extensively variable expression levels measured by high-throughput assays. In this study, we hypothesize that in individual tumor samples, the disruption of transcription homeostasis in key pathways or gene set plays an important role in tumorigenesis and has profound implications for the patient's clinical outcome. We devised a computational method named iPath to identify, at the individual sample level, which pathways or gene sets significantly deviate from their norms.

### Microbial Risk Score for Capturing Microbial Characteristics, Integrating Multi-Omics Data, and Predicting Disease Risk

*♦Chan Wang, Leopoldo Segal, Jiyuan Hu, Boyan Zhou, Richard Hayes, Jiyoung Ahn and Hulin Li*

New York University Grossman School of Medicine  
Chan.Wang@nyulangone.org

## Abstracts

---

With the rapid accumulation of microbiome-wide association studies, a great amount of microbiome data are available to study the microbiome's role in human disease and advance the microbiome's potential use for disease prediction. However, the unique features of microbiome data hinder its utility for disease prediction. Motivated from the polygenic risk score framework, we propose a microbial risk score (MRS) framework to aggregate the complicated microbial profile into a summarized risk score that can be used to measure and predict disease susceptibility. Specifically, the MRS algorithm involves two steps: (1) identifying a sub-community consisting of the signature microbial taxa associated with disease and (2) integrating the identified microbial taxa into a continuous score. Moreover, we propose a multi-omics data integration method by jointly modeling the proposed MRS and other risk scores constructed from other omics data in disease prediction. Through three comprehensive real-data analyses using the NYU Langone Health COVID-19 cohort, the gut microbiome health index (GMHI) multi-study cohort, and a large type 1 diabetes cohort separately, we exhibit and evaluate the utility of the proposed MRS framework for disease prediction and multi-omics data integration.

### **Microbiome Composition-on-Composition Regression Analysis**

Xiang Zhan

Peking University  
zhanx@bjmu.edu.cn

It is quite common to encounter compositional data in a regression framework in data analysis. When both responses and predictors are compositional, most existing models rely on a family of log-ratio based transformations to move the analysis from the simplex to the reals. This often makes interpretation of the model more complex. Motivated by data analysis with microbiome compositional data, a novel transformation-free regression model is proposed that allows for two (or more) compositional predictors to be used via a latent variable mixture. A modified Expectation-Maximization algorithm is proposed to estimate model parameters, which are shown to have natural interpretations. The resulting methodology is evaluated with both numerical simulations and real data applications to demonstrate its validity and superiority.

## **Session 23CHI6: Case Studies in Clinical Trial Design, Analysis and Result Interpretation**

### **Leverage Registry Database and Natural History Study to Facilitate Full Approval—a Real Case Study**

Grace Lin

Sanofi  
Grace1.lin@sanofi.com

Leveraging real-world data (RWD), such as registry database and natural history studies, to support regulatory decision-making is becoming a hot topic. This presentation will introduce a real case in Fabrazyme sBLA using RWD to generate real-world evidence (RWE) to support the full approval. Fabry disease is a rare, X-linked, genetic disorder caused by mutations in the lysosomal enzyme alpha-galactosidase A (GAL) and can potentially impact multiple organs including renal, cerebrovascular, and cardiovascular complications which contribute to major life-threatening end-organ dysfunction. As a result, the most direct and reasonable therapeutic approach is to replace the deficient enzyme activity. Fabrazyme works as an enzyme replacement therapy and has been shown to reduce serum and renal globotriaosylceramide (GL-3), which is regarded as surrogate endpoint to predict clinical benefit of Fab-

razyme. Fabrazyme got approval in the European Union and United States about 20 years ago and cumulated lot of real world data. Full approval of Fabrazyme and extended Label was granted by leveraging prospective observational Fabry Registry and untreated patients from the Natural History study to support the overall effectiveness and long-term clinical benefit of Fabrazyme in the real-world. The whole story of Fabrazyme using RWD will be briefly described in the presentation.

### **Propensity Score Matched External Control: a Use Case in Rare Disease Pediatric Clinical Trial Study**

Ning Li

Sanofi  
liningmacau@gmail.com

In recent years, drug development in rare diseases and pediatric diseases attracts the increased public attention. Pharmaceutical company starts to invest more resources and money into research and development of drug in these areas. By the nature of rare disease and pediatric disease, the population size of patients is not large to begin with and the recruitment of trial participants is difficult. In certain rare diseases, a clinical trial with a single treatment group without placebo control becomes acceptable. This presentation will introduce a rare disease pediatric single-arm clinical trial, and then illustrate how the external registry data is used to construct external control arm using propensity score matching method.

### **Assessing the Probability of Clinical Trial Success via Modeling and Simulation: a Case Study**

Michael Lee

Harbour BioMed, Inc.  
michael.lee@harbourbiomed.com

Drug development is a high-risk investment. To mitigate the risk, continuous assessment of probability of success is common during the product development process. In this presentation we will share an assessment method, which utilizes model-based meta-analysis (MBMA) and simulation. The assessment incorporates all available information, both internal and external data in the assessment. The assessment can start as early as prior to the first clinical trial and continue through the end of pivotal studies. We will use an example to illustrate the assessment method.

## **Session 23CHI71: Addressing Challenges in Time-to-Event Data**

### **Learning Optimal Individualized Treatment Rules with Interval-Censored Data**

Yichi Zhang<sup>1</sup> and ♦Yinghao Pan<sup>2</sup>

<sup>1</sup>Yale University

<sup>2</sup>University of North Carolina at Charlotte  
ypan8@uncc.edu

One central task in personalized medicine is to develop optimal individualized treatment rules that provide personalized treatment recommendations for patients based on their unique characteristics. However, in many clinical studies, the clinical outcome is interval-censored, i.e., the event time is only known to fall within some interval rather than being exactly observed, which complicates the analysis. We propose an ensemble tree-based weighted learning approach for estimating individualized treatment rules with interval-censored data. Our proposed estimator is robust as we do not specify parametric or semi-parametric models for event and censoring times. In simulation studies, our estimators demonstrate improved performance compared to existing methods. Data from human immunod-

efficiency virus (HIV) vaccine/prevention trials will be analyzed as an illustrative example.

### A Semiparametric Approach to Develop Well-Calibrated Risk Assessment Models in Calculating Lifetime Risk for Breast Cancer

♦Yaqi Cao<sup>1</sup>, Ying Yang<sup>2</sup> and Jinbo Chen<sup>3</sup>

<sup>1</sup>Minzu University of China

<sup>2</sup>Tsinghua University

<sup>3</sup>University of Pennsylvania  
yaqicaostats@163.com

The added value of candidate predictors for risk modeling is routinely evaluated by comparing the performance of models with or without including candidate predictors. Such comparison is most meaningful when the estimated risk by the two models are both unbiased in the target population. Oftentimes, data for standard predictors in the base model is richly available from the target population, but data for candidate predictors are available only from nonrepresentative convenience samples. While the base model can be naively updated using the study data without recognizing the discrepancy between the underlying distribution of the study data and that in the target population, the resultant risk estimates as well as evaluation of the candidate predictors are biased. Towards building a well-calibrated updated model, we propose a semiparametric method for model fitting that enforces calibration against a well-calibrated base model. Our method allows unbiased assessment of model improvement by candidate predictors without requiring a representative sample from the target population, thereby overcoming a major bottleneck in practice. Finally, we apply the proposed method to data extracted from Penn Medicine Biobank and develop risk assessment tools to calculate lifetime risk for breast cancer in the Caucasian woman population.

### Cox Model with Left-Truncation and Auxiliary Outcomes

♦Yidan Shi and Sharon X. Xie

University of Pennsylvania Perelman School of Medicine  
yidan.shi@pennmedicine.upenn.edu

The time-to-event analysis is one of the most popular tools for modeling disease process data. When monitoring the time to a disease (e.g., Alzheimer's disease), the biological-based assessment (e.g., cerebrospinal fluid measures of amyloid beta) may be expensive/invasive to measure and is only available for a small group of patients, which brings limitations in sample size and estimating efficiency. In many studies, an auxiliary, inexpensive, or less invasive measure of the disease status can be obtained from a larger sample, such as a diagnosis based on clinical symptoms. We propose a likelihood-based method and an EM algorithm for Cox regression models which incorporate the auxiliary outcomes and improve efficiency. The proposed method allows left-truncation in the event time of interest. We evaluate the proposed method through simulation studies. We illustrate the proposed method using the Alzheimer's Disease Neuroimaging Initiative (ADNI) study.

### Boosting Method for Length-Biased and Interval-Censored Survival Data Subject to High-Dimensional Error-Prone Covariates

♦Li-Pang Chen and Bangxu Qiu

National Chengchi University  
lchen723@nccu.edu.tw

Analysis of length-biased and interval-censored data is an important topic in survival analysis, and many methods have been developed to address this complex data structure. However, these methods focus on low-dimensional data and assume the covariates to be pre-

cisely measured, while high-dimensional data subject to measurement error are frequently collected in applications. We explore a valid inference method for handling high-dimensional length-biased and interval-censored survival data with measurement error in covariates under the accelerated failure time model. We primarily employ the SIMEX method to correct for measurement error effects and propose the boosting procedure to do variable selection and estimation. The proposed method is able to handle the case that the dimension of covariates is larger than the sample size and enjoys appealing features that the distributions of the covariates are left unspecified.

### Session 23CHI72: How to Dissect and Understand Diverse High-Throughput Data by Novel Statistical and Computational Methods?

#### Deconvolution of Bulk RNA-Seq Reveals Cell-Type Specificity Mechanism in Alzheimer's Disease

Yun Li

University of North Carolina at Chapel Hill  
yun\_li@med.unc.edu

Bulk tissue transcriptomic profiles cannot reflect the functional heterogeneity across cell types. We here propose a new empirical Bayes method (EPIC-unmix), that integrates sc/snRNA-seq reference and bulk RNA-seq data from target samples to enhance cell-type-specific (CTS) expression inference in target samples. We applied EPIC-unmix to deconvolute ROSEMAP bulk RNA-seq data from prefrontal cortex samples. Downstream analysis of CTS gene expression, including identification of CTS differentially expressed (DE) genes, CTS eQTL analysis and functional annotations in the corresponding cell type for DE genes, suggested IKZF1 as a risk gene for Alzheimer's disease functioning in microglia.

#### A Prism Vote Method for Risk Prediction of Traits in Genotype Data of Multi-Population

♦Xiaoxuan Xia, Ming Hin Ng, Lin Hou and Yingying Wei  
xia\_xiaoxuan@outlook.com

Multi-population cohorts offer unprecedented opportunities for profiling disease risk in large samples, however, heterogeneous risk effects underlying complex traits across populations make integrative prediction challenging. In this study, we propose a novel Bayesian probability framework, the Prism Vote (PV), to construct risk predictions in heterogeneous genetic data. The PV views the trait of an individual as a composite risk from subpopulations, in which stratum-specific predictors can be formed in data of more homogeneous genetic structure. Since each individual is described by a composition of subpopulation memberships, the framework enables individualized risk characterization. Simulations demonstrated that the PV framework applied with alternative prediction methods significantly improved prediction accuracy in mixed and admixed populations. The advantage of PV enlarges as genetic heterogeneity and sample size increase. In two real genome-wide association data consists of multiple populations, we showed that the framework considerably enhanced prediction accuracy of the linear mixed model in five-group cross validations. The proposed method offers a new aspect to analyze individual's disease risk and improve accuracy for predicting complex traits in genotype data.

#### Identification of Cell-Type-Specific Spatially Variable Genes Accounting for Excess Zeros

Xiangyu Luo

Renmin University of China

## Abstracts

---

xiangyuluo@ruc.edu.cn

Spatial transcriptomic techniques can profile gene expressions while retaining the spatial information, thus offering unprecedented opportunities to explore the relationship between gene expression and spatial locations. The spatial relationship may vary across cell types, but there is a lack of statistical methods to identify cell-type-specific spatially variable (SV) genes by simultaneously modeling excess zeros and cell-type proportions. Results We develop a statistical approach CTSV to detect cell-type-specific SV genes. CTSV directly models spatial raw count data and considers zero-inflation as well as overdispersion using a zero-inflated negative binomial distribution. It then incorporates cell-type proportions and spatial effect functions in the zero-inflated negative binomial regression framework. The R package pscl is employed to fit the model. For robustness, a Cauchy combination rule is applied to integrate P-values from multiple choices of spatial effect functions. Simulation studies show that CTSV not only outperforms competing methods at the aggregated level but also achieves more power at the cell-type level. By analyzing pancreatic ductal adenocarcinoma spatial transcriptomic data, SV genes identified by CTSV reveal biological insights at the cell-type level.

### Gene-Environment Interaction Analysis via Deep Learning

Shuni Wu<sup>1</sup>, Yaqing Xu<sup>2</sup>, Qingzhao Zhang<sup>3</sup> and ♦Shuangge Ma<sup>4</sup>

<sup>1</sup>The Wang Yanan Institute for Studies in Economics, Xiamen University

<sup>2</sup>Shanghai Jiao Tong University School of Medicine

<sup>3</sup>Department of Statistics and Data Science, School of Economics and Fujian Key Lab of Statistics, Xiamen University

<sup>4</sup>Department of Biostatistics, Yale School of Public Health  
yaqing.xu@aya.yale.edu

Gene-environment (G-E) interaction analysis plays an important role in studying complex diseases. Extensive methodological research has been conducted on G-E interaction analysis, and the existing methods are mostly based on regression techniques. In many fields including biomedicine and omics, it has been increasingly recognized that deep learning may outperform regression with its unique flexibility (for example, in accommodating unspecified non-linear effects) and superior prediction performance. However, there has been a lack of development in deep learning for G-E interaction analysis. In this article, we fill this important knowledge gap and develop a new analysis approach based on deep neural network in conjunction with penalization. The proposed approach can simultaneously conduct model estimation and selection (of important main G effects and G-E interactions), while uniquely respecting the "main effects, interactions" variable selection hierarchy. Simulation shows that it has superior prediction and feature selection performance. The analysis of data on lung adenocarcinoma (LUAD) and skin cutaneous melanoma (SKCM) overall survival further establishes its practical utility. Overall, this study can advance G-E interaction analysis by delivering a powerful new analysis approach based on modern deep learning.

## Session 23CHI74: Recent Advances in Functional Data Analysis

### A Unified Analysis of Multi-Task Functional Linear Regression Models with Manifold Constraint and Composite Quadratic Penalty

Shiyuan He

Renmin University

heshiyuan@ruc.edu.cn

This work studies the multi-task functional linear regression models where both the covariates and the unknown regression coefficients (called slope functions) are curves. For slope function estimation, we employ penalized splines to balance bias, variance, and computational complexity. The power of multi-task learning is brought in by imposing additional structures over the slope functions. We propose a general model with double regularization over the spline coefficient matrix: i) a matrix manifold constraint, and ii) a composite penalty as a summation of quadratic terms. Many multi-task learning approaches can be treated as special cases of this proposed model, such as a reduced-rank model and a graph Laplacian regularized model. We show the composite penalty induces a specific norm, which helps to quantify the manifold curvature and determine the corresponding proper subset in the manifold tangent space. The complexity of tangent space subset is then bridged to the complexity of geodesic neighbor via generic chaining. A unified convergence upper bound is obtained and specifically applied to the reduced-rank model and the graph Laplacian regularized model. The phase transition behaviors for the estimators are examined as we vary the configurations of model parameters.

### Exponential-Family Principal Component Analysis of Two-Dimensional Functional Data with Serial Correlation

Shirun Shen<sup>1</sup>, ♦Kejun He<sup>1</sup>, Bohai Zhang<sup>2</sup> and Lan Zhou<sup>3</sup>

<sup>1</sup>Renmin University of China

<sup>2</sup>Nankai University

<sup>3</sup>Texas A&M University  
kejunhe@ruc.edu.cn

Motivated by a study on Arctic sea-ice-extent (SIE) data of binary observations, in this paper, we propose a novel model to analyze serially correlated non-Gaussian data observed on a two-dimensional domain that may have an irregular shape. We assume the observed data follow a distribution from the exponential family, where the corresponding natural parameter is a dynamic smooth function of two-dimensional locations. A functional principal component model using bivariate splines defined on triangulations is applied on the natural-parameter surface to characterize the spatial variation of data. The serial correlation of data observed at consecutive time points is modeled by autoregressive (AR) processes on the principal component scores. To estimate the unknown parameters, we develop an EM algorithm with two approaches, using Laplace approximation and variational inference, respectively, on the E-step. Through simulation studies, we find that the latter is much faster with higher estimation accuracy, especially when the sample size is large. Finally, the proposed model with variational inference EM algorithm is applied to analyze the massive monthly Arctic SIE data.

### Latent Group Detection in Functional Partially Linear Regression Models

♦Wu Wang<sup>1</sup>, Ying Sun<sup>2</sup> and Huixia Wang<sup>3</sup>

<sup>1</sup>Center for Applied Statistics and School of Statistics, Renmin University of China

<sup>2</sup>Statistics Program, King Abdullah University of Science and Technology

<sup>3</sup>Department of Statistics, The George Washington University, Washington  
wu.wang@ruc.edu.cn

In this paper, we propose a functional partially linear regression model with latent group structures to accommodate the heterogeneous relationship between a scalar response and functional covariates. The proposed model is motivated by a salinity tolerance study

of barley families, whose main objective is to detect salinity tolerant barley plants. Our model is flexible, allowing for heterogeneous functional coefficients while being efficient by pooling information within a group for estimation. We develop an algorithm in the spirit of K-means clustering to identify latent groups of the subjects under study. We establish the consistency of the proposed estimator, derive the convergence rate and the asymptotic distribution, and develop inference procedures. We show by simulation studies that the proposed method has higher accuracy for recovering latent groups and for estimating the functional coefficients than existing methods. The analysis of the barley data shows that the proposed method can help identify groups of barley families with different salinity tolerant abilities.

### Global Depths for Irregularly Observed Multivariate Functional Data

Zhuo Qu<sup>1</sup>, ♦Wenlin Dai<sup>2</sup> and Marc G. Genton<sup>1</sup>

<sup>1</sup>KAUST

<sup>2</sup>Renmin University of China  
wenlin.dai@ruc.edu.cn

Two frameworks for multivariate functional depth based on multivariate depths are introduced in this paper. The first framework is multivariate functional integrated depth, and the second framework involves multivariate functional extremal depth, which is an extension of the extremal depth for univariate functional data. In each framework, global and local multivariate functional depths are proposed. The properties of population multivariate functional depths and consistency of finite sample depths to their population versions are established. In addition, finite sample depths under irregularly observed time grids are estimated. As a by-product, the simplified sparse functional boxplot and simplified intensity sparse functional boxplot are proposed for visualization without data reconstruction. A simulation study demonstrates the advantages of global multivariate functional depths over local multivariate functional depths in outlier detection and running time for big functional data. An application of our frameworks to cyclone tracks data demonstrates the excellent performance of our global multivariate functional depths.

### Session 23CHI75: Recent Statistical Advances for Complex Multi-Omics Data Analysis

#### Kernel-Based Multi-Omics Data Integration for Cancer Subtyping

Hongyan Cao

Shanxi Medical University  
caohy@sxmu.edu.cn

Differentiating cancer subtypes is crucial to guide personalized treatment and improve the prognosis for patients. Integrating multi-omics data can offer a comprehensive landscape of cancer biological process and provide promising ways for cancer diagnosis and treatment. Toward this goal, many multi-omics data integration methods have been developed. Among them, kernel-based methods (e.g. SNF, CIMLR) taking advantage of the ‘kernel trick’ have been widely applied to cancer subtype identification. This approach uses a predefined set of kernels to combine multi-omics data, which avoids the potential issue brought up by feature pre-screening. However, both SNF and CIMLR have some limitations. For SNF, the unavoidable noise features generated by the limitations of measurement technology and inherent natural variation, may dilute clustering signals and lead to some spurious associations between samples. For CIMLR, it optimizes kernel parameters of

all data types and weight of each kernel simultaneously, thus may ignore the heterogeneity of different omics data types and further weaken the clustering power. To alleviate these disadvantages, we developed three multi-omics data integration methods. Based on SNF, we incorporated network enhancement strategy to denoise the similarity network fusion (ne-SNF), and we also proposed a joint similarity network fusion (Joint-SNF) method. Based on CIMLR, we

#### Multi-Omics Data Integration with Multi-View Learning via Composed Tensors

♦Xu Liu<sup>1</sup>, Yiming Lliu<sup>1</sup>, Xiao Zhang<sup>1</sup>, Jian Huang<sup>2</sup>, Yuehua Cui<sup>3</sup> and Xingjie Shi<sup>4</sup>

<sup>1</sup>Shanghai University of Finance and Economics

<sup>2</sup>The Hong Kong Polytechnic University

<sup>3</sup>Michigan State University

<sup>4</sup>East China Normal University  
liu.xu@sufe.edu.cn

The integration of multi-views and multiple outcomes enables comprehensive quantification of the underlying biological mechanisms of complex diseases. To capture non-linear behavior of a complex biological system, it is natural to model non-linear relationships with an additive regression model with each component modelled nonparametrically. The need of estimating a large number of component functions given limited data poses a great challenge, owing to the large number of parameters. To overcome the 20 challenge, we propose a sparse tensor reduced rank approach to a multivariate additive regression for multi-view (MARM). The B-spline coefficients in each view are treated as a third-order tensor with low rankness. It dramatically reduces the number of free parameters. Since some views are measurements of different layers for shared basic units, there is consensus information among these views. To leverage this consensus information, we further propose a composed MARM (COMARM) that rearranges the spline 25 coefficients from those views for the shared units into a fourth-order tensor. The Tucker decomposition with a group-sparse penalty is applied to encourage the agreement of output of different views and yield a sparse and low-rank tensor estimator. To efficiently fit the model, we design an alternative updating algorithm based on

#### Leveraging Trans-Ethnic Genetic Risk Scores to Improve Association Power for Complex Traits in Underrepresented Populations

Haojie Lu, Shuo Zhang, Zhou Jiang and ♦Ping Zeng

Xuzhou Medical University  
zpsstat@zjmu.edu.cn

Trans-ethnic genome-wide association studies have revealed that many loci identified in European populations can be reproducible in non-European populations, indicating widespread trans-ethnic genetic similarity. However, how to leverage such shared information more efficiently in association analysis is less investigated for traits in underrepresented populations. We here propose a statistical framework, trans-ethnic genetic risk score informed gene-based association mixed model (GAMM), by hierarchically modeling SNP effects in the target population as a function of effects of the same trait in well-studied populations. GAMM powerfully integrates genetic similarity across distinct ancestral groups to enhance power in understudied populations, as confirmed by extensive simulations. We illustrate the usefulness of GAMM via the application to thirteen blood cell traits in Africans of the UK Biobank ( $n=3,204$ ) while utilizing genetic overlap shared in Europeans ( $n=746,667$ ) and East Asians ( $n=162,255$ ). We discovered multiple new asso-

## Abstracts

---

ciated genes which had otherwise been missed by existing methods, and revealed the trans-ethnic information indirectly contributed much to the phenotypic variance. Overall, GAMM represents a flexible and powerful statistical framework of association analysis for complex traits in underrepresented populations by integrating trans-ethnic genetic similarity across well-studied populations, and helps attenuate health inequities in current genetics research for people of minority populations.

### An Adaptive Set-Based Testing Framework for High-Dimensional Association Studies

Haitao Yang

haitaoyang@hebmu.edu.cn

Set-based association analysis that jointly tests the association of variants in a region or group has been a powerful tool in understanding the etiology of complex diseases in GWAS studies. When evaluating a SNP-set association, two types of disease-SNP functional model are commonly recognized: 1) there are multiple functional variants each with small effect in a set, and together they contribute to a disease risk. We term this model as the cumulative weak signal model (CWSM); and 2) Very few functional variants each with large dominating effect in a set contribute to a disease risk. We term this model as the dominating strong signal model (DSSM). There are two main issues limit the power of existing methods: 1) existing methods work only under either CWSM or DSSM. Any misspecification of the disease model can substantially lead to power loss; and 2) the high-dimensional nature of SNP variants is typically not considered, resulting in low power or high false positives. In this work, we adopt a high-dimensional inference procedure when fitting many SNPs in a regression model simultaneously. Then, we propose an omnibus testing method with a robust and powerful p-value combination method to boost the power of SNP-set association.

### Session 23CHI80: Some Advances in Analysis of High-Dimensional Complex Data

#### Distribution Estimation of Contaminated Data via Dnn-Based Mom-Gans

Fang Xie<sup>1</sup>, Lihu Xu<sup>2</sup>, Qiuran Yao<sup>2</sup> and ♦Huiming Zhang<sup>3</sup>

<sup>1</sup>BNU-HKBU United International College

<sup>2</sup>University of Macau

<sup>3</sup>Beihang University  
zhanghuiming@buaa.edu.cn

The traditional adversarial nets, for example, the generative adversarial network (GAN), are sensitive to contaminated data. In this talk, we develop a robust and deep neural network (DNN) method to estimate the learning distribution based on the adversarial nets and median-of-mean (MoM) approach, and it is called the MoM-GAN method. Theoretically, we obtain a non-asymptotic error bound for the DNN-based Wasserstein-1 MoM-GANs estimator measured by integral probability metrics with the H

"older function class. The derived finite-sample high-probability error-bounds concern the outlier proportion and the fraction of sane blocks. We give an algorithm for our proposed method and implement it through two real applications, which show that our proposed method outperforms Wasserstein GAN for contaminated data.

#### Large-Scale Spatial Multiple Testing via Shifted p-Values

Pengyu Yan and ♦Pengfei Wang

Dongbei University of Finance and Economics  
wangpf0429@dufe.edu.cn

Large-scale multiple testing, namely, simultaneous testing of tens of thousands of hypotheses, has been widely employed in many scientific fields. Conventional multi- ple testing procedures usually assume that different hypotheses are exchangeable and broadly ignore the spatial information. However, in many practical applications, the neighbouring tests exhibit complex correlations so that the corresponding hypotheses are not exchangeable. In addition, the proper use of the spatial information is ex- pected to improve both the power of multiple testing and the interpretability of the findings. This paper develops a novel large-scale spatial multiple testing procedure based on the shifted p-value, which carries two types of information, one is the infor- mation from the signal and the other is the information from the location-adaptive sparsity level. Theoretical results show that the proposed multiple testing procedure is valid, namely, it is capable of controlling the false discovery rate (FDR) at the pre- specified level. An expectation-maximization (EM) is developed for estimating the location-adaptive sparsity level and the turning parameter. The simulations and re- al data analysis demonstrate the superiority of the proposed procedure in large-scale spatial multiple testing.

#### Adaptive Learning of Personalized Tuning Parameters for Feature Selection in Linear Models

Bin Wang, ♦Xiaofei Wang and Jianhua Guo

Northeast Normal University

wangxf341@nenu.edu.cn

Feature selection in linear models is widely applied to diverse fields. Many traditional approaches can behave well based on the penalized likelihood. However, the selection of hyperparameters or tuning parameters is often critical and challenging. From the oracle viewpoint for attaining an ideal risk, we design a global adaptive generative adjustment algorithm, which can adaptively learn multiple tuning parameters in the personalized Tikhonov regularization and further select features by a personalized thresholding strategy. Finally, in the numerical experiment, we compared our algorithms to the LASSO, the SCAD, and the MCP. The experiment results affirmed the efficiency of our algorithms for feature selection and demonstrated the superiority of our algorithms over other methods.

#### On Statistical Analysis of High-Dimensional Factor Models

Junfan Mao, ♦Zhigen Gao, Bingyi Jing and Jianhua Guo

gaoz112@nenu.edu.cn

High-dimensional factor models have received much attention since the seminal work of Bai and Li (2012). We make several contributions to the asymptotic properties of the quasi maximum likelihood estimates(MLEs) as both the sample size T and the variable dimension N go to infinity. First we eliminate one of rather unnatural assumptions on the variance estimates which is commonly assumed in the literature. Secondly, we give unified results on the asymptotic properties of the quasi MLEs, which greatly expand the scope of earlier studies. Simulations are given to illustrate these results.

### Session 23CHI83: New Advances in Statistical Methods for Health-Related Data

#### Admissibility Condition for Combining the Dependent p-Values and Its Application for Meta-Analysis

♦Ke Yang<sup>1</sup> and Tiejun Tong<sup>2</sup>

<sup>1</sup>Beijing University of Technology

<sup>2</sup>Hong Kong Baptist University  
yangke@bjut.edu.cn

Combining the  $p$ -values is an important statistical approach with applications in signal detection, meta-analysis, etc. Existing methods and their statistical properties for combining the  $p$ -values rely on the assumption that the individual  $p$ -values are independent of each other. In this paper, we propose new methods that are able to combine the  $p$ -values derived from dependent tests. For this purpose, we first derive the joint distribution of the bivariate  $p$ -values that are derived from testing two dependent normal means. Like Birnbaum's admissibility for methods of combining the independent  $p$ -values, we also propose the criterion of admissibility for combining the dependent  $p$ -values and derive the set of methods combining the bivariate  $p$ -values that satisfy such criterion. The theoretical results for the bivariate  $p$ -values are further generalized to the multivariate  $p$ -values. It is shown that for dependent case, Stouffer's combination method is also admissible subject to an adjustment on the critical values. By simulations and an application to meta-analysis, we show that the adjusted Stouffer's combination method can achieve its nominal type I error rate more accurately than the unadjusted competitor.

### **Standardized Mean Difference without the Homoscedasticity Assumption**

♦*Jiandong Shi<sup>1</sup>, Xiaochen Zhang<sup>2</sup> and Tiejun Tong<sup>3</sup>*

<sup>1</sup>The Hong Kong University of Science and Technology

<sup>2</sup>Shandong University

<sup>3</sup>Hong Kong Baptist University  
eejiandong@ust.hk

In the case and control studies for the continuous outcomes, the standardized mean difference (SMD) is a commonly used effect size to quantify the difference between the case and control groups. The SMD is usually defined with the homoscedasticity assumption, which may be easily violated in practice. Despite that the heteroscedasticity case is also considered in the literature and several related definitions are given, they have the respective disadvantages as we point out. In this paper, we propose a new definition of SMD without the homoscedasticity assumption, and provide its unbiased estimator and the confidence interval. Simulation studies are carried out to demonstrate that our proposed estimator as well as the confidence interval performs well in terms of relative bias, relative mean square error and coverage probability, for both homoscedasticity and heteroscedasticity cases. We also perform a real data analysis to illustrate its usefulness in meta-analysis.

### **Sparse Convolved Rank Regression in High Dimensions**

♦*Le Zhou<sup>1</sup>, Boxiang Wang<sup>2</sup> and Hui Zou<sup>3</sup>*

<sup>1</sup>Hong Kong Baptist University

<sup>2</sup>University of Iowa

<sup>3</sup>University of Minnesota  
lezhou@hkbu.edu.hk

Wang et al. (2020, JASA) studied the high-dimensional sparse penalized rank regression and established its nice theoretical properties. Compared with the least squares, rank regression can have a substantial gain in estimation efficiency while maintaining a minimal relative efficiency of 86.4%. However, the computation of penalized rank regression can be very challenging for high-dimensional data, due to the highly nonsmooth rank regression loss. In this work we view the rank regression loss as a non-smooth empirical counterpart of a population level quantity, and a smooth empirical counterpart is derived by substituting a kernel density estimator for the true distribution in the expectation calculation. This view leads to the convoluted rank regression loss and consequently the sparse penalized convoluted rank regression (CRR) for high-

dimensional data. We prove some interesting asymptotic properties of CRR. Under the same key assumptions for sparse rank regression, we establish the rate of convergence of the  $\ell_1$ -penalized CRR for a tuning free penalization parameter and prove the strong oracle property of the folded concave penalized CRR. We further propose a high-dimensional Bayesian information criterion for selecting the penalization parameter in folded concave penalized CRR and prove its selection consistency. We derive an efficient algorithm for solving sparse

### **A Nonparametric Mixed-Effects Mixture Model for Patterns of Clinical Measurements Associated with Covid-19**

*Xiaoran Ma<sup>1</sup>, Wensheng Guo<sup>2</sup>, Mengyang Gu<sup>1</sup>, Peter Kotanko<sup>3</sup>, Len Usvyat<sup>4</sup> and ♦Yuedong Wang<sup>1</sup>*

<sup>1</sup>University of California - Santa Barbara

<sup>2</sup>University of Pennsylvania

<sup>3</sup>Renal Research Institute

<sup>4</sup>Fresenius Medical Care  
yuedong@ucsb.edu

Some patients with COVID-19 show changes in signs and symptoms, such as temperature and oxygen saturation, days before being positively tested for SARS-CoV-2, while others remain asymptomatic. It is important to identify these subgroups and to understand what biological and clinical predictors are related to these subgroups. This information will provide insights into how the immune system may respond differently to infection and can further be used to identify infected individuals. We propose a flexible nonparametric mixed-effects mixture model that identifies risk factors and classifies patients with biological changes. We model the latent probability of biological changes using a logistic regression model and trajectories in the latent groups using smoothing splines. We developed an EM algorithm to maximize the penalized likelihood for estimating all parameters and mean functions. We evaluate our methods by simulations and apply the proposed model to investigate changes in temperature in a cohort of COVID-19-infected hemodialysis patients.

### **Session 23CHI95: Modern Statistical Process Control and Change-Point Problems I**

#### **Design of Ewma Control Chart with Time-Varying Smoothing Parameters Based on Bayesian Likelihood Ratio**

♦*Baobai Guo and Pingye Gong*

Zhejiang Gongshang University  
gbc78@163.com

This paper proposes a control chart with time-varying smoothing parameter based on Bayesian generalized likelihood ratio test for monitoring the process mean. For the design of time-varying smoothing parameter, in addition to considering the previously proposed time-varying smooth parameters, this article also proposes a new time-varying smoothing parameter based on the posterior predictive distribution of the mean  $\hat{\mu}$ . Compared with the previously proposed time-varying smoothing parameters, it is more convenient to not require any parameters to be specified. The compared results show that the proposed chart performs better than the existing charts.

#### **A New Ewma Control Chart for Monitoring Variability**

*Dan Wang*

Northwest University  
wangdan@nwu.edu.cn

## Abstracts

---

Most of the suggested nonparametric control charts are applied to monitor the location parameter. In this paper, we propose a nonparametric exponentially weighted moving average(EWMA) control chart for monitoring scale parameter, which combines Ansari-Bradley test and the framework of change point detection, start-up control charts do not require the assumption that the prior distribution of the process is known. For most nonparametric control charts takes a large amount of historical observations to motivate the control chart, our proposed nonparametric EWMA control chart achieve the same purpose with a very small number of historical observations. Simulation results prove that our control chart is robust in monitoring the scale parameter shift, and finally, an application example demonstrates the practicality of this control chart for monitoring the parameter of scale.

### **Fault Diagnosis for High-Dimensional Data Streams under Two-Layers Mdr Controls**

*Dongdong Xiang*

East China Normal Univesity  
terryxdd@163.com

This paper is concerned with fault diagnosis for high dimensional data streams when there are multiple points in each line with potential points of fault. The proposed method utilizes a compound decision theoretical framework with multiple testing to develop a decision rule that simultaneously controls two layers of missing discovery rates while minimizing the expected false positive rate. Numerical analysis was performed using both oracle and data-driven procedures, comparing the results with existing diagnostic procedure. The analysis demonstrated that the proposed method outperformed its competitors. Application to real-world semiconductor manufacturing data illustrated its effectiveness in identifying faults in complex data streams.

### **On – line Profile Monitoring for Surgical Outcomes using a Weighted Score Test**

*♦Liu Liu, Xin Lai, Jian Zhang and Fugee Tsung*  
liuliu@sicnu.edu.cn

In the past decade, risk-adjusted control charts have been widely used to monitor surgical outcomes. However, most existing approaches focus on monitoring shifts in location parameters and may not be able to detect a scale change that is also likely to occur in surgical data. We propose a new charting method to simultaneously monitor location and scale parameters under the generalized linear mixed model (GLMM) framework. We derive a weighted score test statistic to construct the exponentially weighted moving average chart. This new chart may be applied to monitor surgical performance. Simulation results indicate that the proposed method is more efficient than existing methods such as the risk-adjusted cumulative sum (RA-CUSUM) chart in detecting the heterogeneity of surgical outcomes. A real data set from the Surgical Outcome Monitoring and Improvement Program in Hong Kong is used to illustrate the applicability of the proposed chart.

## **Session 23CHI107: New Developments of Statistical Methods in Bio-Medical Studies**

### **Estimation of Quantile Function for Gene Expression Trajectory under Multiple Biological Conditions**

*♦Dianliang Deng and Qi Lyu*  
University of Regina  
deng@uregina.ca

Containing rich information to characterize gene function, temporal gene expression data is widely applied to biomedical studies. Gene expression trajectories under different biological conditions are mostly analyzed through traditional mean regression model, which usually does not work in practice because of the non-normal distribution and heteroscedasticity of the data. This research proposes a likelihood-based EM algorithm to estimate marginal conditional quantile of multivariate response. We assume the error term follows multivariate asymmetric Laplace distribution, and implement the EM algorithm based on MLE calculations. The proposed approach is first validated through simulation studies, then it is utilized to analyze a real dataset including 18 genes in *P. aeruginosa* expressed under 24 biological conditions.

### **New Developments of Statistical Methods in Bio-Medical Studies**

*♦Larry Tang and Ty Nguyen*

University of Central Florida  
liansheng.tang@ucf.edu

In medical imaging studies with ordinal ratings, the underlying receiver operating characteristic (ROC) curve is commonly estimated using an ordinal regression model. In this talk, I will discuss a homogeneity test for covariate-adjusted ROC curves based on the ordinal regression. Moreover, pairwise comparison tests are also investigated by establishing confidence intervals of difference among the area under the curves and confidence bands of difference among ROC curves. In addition, the relationship between sample sizes and the power of the proposed homogeneity test is examined to determine the minimum required number of samples.

### **Identification of Survival Relevant Genes with Measurement Error in Gene Expression Incorporated**

*Juan Xiong<sup>1</sup> and ♦Wenqing He<sup>2</sup>*

<sup>1</sup>Shengzhen University

<sup>2</sup>University of Western Ontario  
whe@stats.uwo.ca

Modern gene expression technologies enable the simultaneous measurement of thousands of genes and thus are important for predicting patient survival. However, survival analysis with gene expression data is challenging due to the high dimensionality. Proper identification of survival-relevant genes is imperative for building suitable prediction models. Gene expressions are typically subject to measurement errors introduced from the complex experimental procedure and the measurement error is often ignored. In this talk, the effect of measurement error on the identification of survival-relevant genes is explored under the accelerated failure time model. Survival-relevant genes are identified by regularizing the weighted least square estimator with the adaptive LASSO penalty. The simulation-extrapolation method is applied to adjust for the impact of measurement error. The performance of the proposed method is assessed by simulation studies and illustrated by a real study.

### **Direct Estimation of Volume under the Roc Surface with Verification Bias**

*♦Gengsheng Qin and Shuangfei Shi*  
gqin@gsu.edu

In practice, the receiver operating characteristic (ROC) curve of a diagnostic test is widely used to show the performance of the test for discriminating two-class events. The area under the ROC curve (AUC) is proposed as an index for the assessment of the diagnostic accuracy of the test under consideration. Due to ethical and cost considerations associated with application of gold standard (GS)

tests, only a subset of the patients initially tested have verified disease status. Statistical evaluation of the test performance based only on test results from subjects with verified disease status are typically biased. Various AUC estimation methods for tests with verification biased data have been developed over the last few decades. In this article, we develop new direct estimation methods for the volume under the ROC surface (VUS) by extending the AUC estimation methods for two-class diagnostic tests to three-class diagnostic tests in the presence of verification bias. The proposed methods will provide a comprehensive guide to deal with the verification bias in three-class diagnostic test accuracy studies and lead to a better choice of diagnostic tests.

## Session 23CHI11: Recent Development on High-Dimensional Data Analysis

### An Adaptable Independence Test using Kernel Projection Criterion in High Dimensions

*Yuexin Chen and ♦Wangli Xu  
wlxu@ruc.edu.cn*

Testing the independence between two high-dimensional random vectors is a fundamental and challenging problem in statistics. Most existing tests based on distance and kernel may fail to detect the non-linear dependence in the high-dimensional regime. To tackle this obstacle, this paper proposes an adaptable kernel independence test for assessing the independence between two random vectors based on a class of Gaussian projections relying on sparsity parameters. The proposed test can be generally implemented for a wide class of distance-based kernels and completely characterizes dependence in the lowdimensional regime. Besides, the test captures pure nonlinear dependence in the high-dimensional regime. Theoretically, we develop central limit theorem and rate of convergence for the proposed statistic under some mild regularity conditions and the null hypothesis. Moreover, we derive the asymptotic power of the proposed test for a certain type of alternative which enables us to select suitable sparsity parameters to achieve superior power in the high-dimensional regime. The adaptive choices of sparsity parameters ensure that the proposed test has comparable power with the original kernel-based test in the moderately high-dimensional regime. Numerical experiments also demonstrate the satisfactory empirical performance of the proposed test in various scenarios.

### Two-Sample Test for High-Dimensional Covariance Matrices: a Normal-Reference Approach

*♦Jin-Ting Zhang<sup>1</sup>, Jingyi Wang<sup>2</sup> and Tianming Zhu<sup>3</sup>*

<sup>1</sup>Professor

<sup>2</sup>PhD student

<sup>3</sup>Assistant Professor  
stazjt2020@nus.edu.sg

Abstract Testing the equality of the covariance matrices of two high-dimensional samples is a fundamental inference problem in statistics. Several tests have been proposed but they are either too liberal or too conservative when the required assumptions are not satisfied. This means that they are not always applicable in real data analysis. To overcome this difficulty, a normal-reference test is proposed and studied in this paper. It is shown that under some regularity conditions and the null hypothesis, the proposed test statistic and a chi-square-type mixture have the same limiting distribution. It is then justified to approximate the null distribution of the proposed test statistic using that of the chi- square-type mixture. The distribu-

tion of the chi-square-type mixture can be well approximated using a three-cumulant matched chi-square approximation with its approximation parameters consistently estimated from the data. The asymptotic power of the proposed test under a local alternative is also established. Simulation studies and a real data example demonstrate that in terms of size control, the proposed test outperforms the existing competitors substantially. KEY WORDS: chi-square-type mixtures; high-dimensional data; three-cumulant matched chi-square approximation; equal covariance matrices.

### A Pairwise Hotelling Method for Testing High-Dimensional Mean Vectors

*♦Zongliang Hu<sup>1</sup>, Tiejun Tong<sup>2</sup> and Marc G. Genton<sup>3</sup>*

<sup>1</sup>Shenzhen University

<sup>2</sup>HongKong Baptist University

<sup>3</sup>King Abdullah University of Science and Technology  
zhu@szu.edu.cn

TBC

## Session 23CHI14: Topics on Statistical Study and Method

### Matrix Garch Model: Inference and Application

*♦Cheng Yu<sup>1</sup>, Dong Li<sup>1</sup>, Feiyu Jiang<sup>2</sup> and Ke Zhu<sup>3</sup>*

<sup>1</sup>Tsinghua University

<sup>2</sup>Fudan University

<sup>3</sup>The University of Hong Kong  
yuc20@mails.tsinghua.edu.cn

Matrix-variate time series data are largely available in applications. However, no attempt has been made to study their conditional heteroskedasticity which is often observed in economic and financial data. To address this gap, we propose a novel matrix generalized autoregressive conditional heteroskedasticity (GARCH) model to capture the dynamics of conditional row and column covariance matrices of matrix time series. The key innovation of the matrix GARCH model is the use of a univariate GARCH specification for the trace of conditional row or column covariance matrix, which allows for the identification of conditional row and column covariance matrices. Moreover, we introduce a quasi-maximum likelihood estimator (QMLE) for model estimation and develop a portmanteau test for model diagnostic checking. Simulation studies are conducted to assess the finite-sample performance of the QMLE and portmanteau tests. To handle large dimensional matrix time series, we also propose a matrix factor GARCH model. Finally, we demonstrate the superiority of the matrix GARCH and matrix factor GARCH models over existing multivariate GARCH-type models in volatility forecasting and portfolio allocations using three applications on credit default swap prices, global stock sector indices, and future prices.

### On Bivariate Threshold Poisson Integer-Valued Autoregressive Processes

*♦Kai Yang<sup>1</sup>, Yiwei Zhao<sup>1</sup>, Han Li<sup>2</sup> and Dehui Wang<sup>3</sup>*

<sup>1</sup>Changchun University of Technology

<sup>2</sup>Changchun University

<sup>3</sup>Liaoning University  
yangkai@ccut.edu.cn

To capture the bivariate count time series showing piecewise phenomena, we introduce a first-order bivariate threshold Poisson integer-valued autoregressive process. Basic probabilistic and statistical properties of the model are discussed. Conditional least squares and conditional maximum likelihood estimators, as well as

## Abstracts

---

their asymptotic properties, are obtained for both the cases that the threshold parameter is known or not. A new algorithm to estimate the threshold parameter of the model is also provided. Moreover, the nonlinearity test and forecasting problems are also addressed. Finally, some numerical results of the estimates and a real data example are presented.

### Semi-Supervised Learning for Two-Sample Comparison

*Mengjiao Peng*

East China Normal University  
mjpeng@fem.ecnu.edu.cn

TBC

## Session 23CHI15: Lifetime Data Analysis

### Group Sequential Trial Design without Proportional Hazards Assumption

*Yiming Chen<sup>1</sup>, John Lawrence<sup>1</sup> and ♦Mei-Ling Ting Lee<sup>2</sup>*

<sup>1</sup>US Food & Drug Administration

<sup>2</sup>University of Maryland  
mltlee@umd.edu

The Non-Proportional Hazard (NPH) phenomenon is not uncommon in medical research. Traditional methods such as Cox model, Log-Rank test lose power significantly when the PH assumption is violated. We use the first hitting-time based Threshold regression (TR) model to address the NPH challenge in survival analysis and generalize the method for group sequential design. We formalize clinical trial design using TR. The hypotheses would be based on the biological mechanism of the treatment effects. Power/sample size formulas are also derived. An application of TR on real clinical data is also presented. We show that, in sequential design, TR is robust under different trends of hazards, it also provides the largest early stopping probabilities comparing to other sequential design which do not require the PH assumption.

### Semiparametric Estimation of Cause-Specific Regression Parameters and Cumulative Incidence Functions for Serial Gap Time Data with Recurrent Events and a Terminal Event

*Shu-Hui Chang*

National Taiwan University  
shuhui@ntu.edu.tw

Serial events including recurrent events and a terminal event are often observed in a longitudinal study of a chronic disease. When recurrence occurs, therapeutic intervention is performed immediately in clinical practice. Therefore, the gap times to an event, recurrence or terminal event from a previous recurrence are the natural outcomes of interest. Semiparametric regression models are introduced to model a sequence of episode-cause-specific hazards for such serial gap times by accounting for cause-specific covariate effects. Two types of generalized linear hazards models for the censoring time are considered to handle dependent censoring induced by covariates and the dependence among serial gap times. We then introduce the inverse probability-of-censoring weights to construct the semiparametric methods for estimating cause-specific covariate effects without specifying the pattern of association between serial gap times. The estimated cumulative incidence functions are also provided. The proposed methods are applied to a thyroid cancer data set for illustration.

## Session 23CHI153: Advances in Regression Methods and Study Design

### Linearized Maximum Rank Correlation Estimation

*Guohao Shen<sup>1</sup>, Kani Chen<sup>2</sup>, Jian Huang<sup>1</sup> and ♦Yuanyuan Lin<sup>3</sup>*

<sup>1</sup>The Hong Kong Polytechnic University

<sup>2</sup>Hong Kong University of Science and Technology

<sup>3</sup>The Chinese University of Hong Kong  
ylin@sta.cuhk.edu.hk

We propose a linearized maximum rank correlation estimator for the single index model. Unlike the existing maximum rank correlation and other rank-based methods, the proposed estimator has a closed-form expression, making it appealing in theory and computation. The proposed estimator is robust to outliers in the response and its construction does not need the knowledge of the unknown link function or the error distribution. Under mild conditions, it is shown to be consistent and asymptotically normal when the predictors satisfy the linearity of expectation assumption. A more general class of estimators is also studied. Inference procedures based on the plug-in rule or random weighting resampling are employed for variance estimation. The proposed method can be easily modified to accommodate censored data. It can also be extended to deal with high-dimensional data combined with a penalty function. Extensive simulation studies provide strong evidence supporting that the proposed method works well in various practical situations.

### Complex Innovative Design Pilot Program and a Potential Proposal: Bayesian Predictive Platform Design for Proof of Concept and Dose Finding using Early and Late Endpoints

*Li Wang*

wangleelee@gmail.com

Complex innovative design (CID) pilot program from FDA provided an excellent opportunity for industry statisticians to think out of the box and to push the boundary on trial designs for late stage studies. We propose a potential Bayesian CID in SLE to combine Bayesian dose ranging, master protocol, group sequential design and short term/long term biomarker together to speed up clinical development and increase the probability of success for the platform. Design parameters including the maximum sample size are calibrated to obtain good frequentist properties such as type I error and power. We use a hypothetical trial of three agents for systemic lupus erythematosus to illustrate the concept and extensive simulations show that the proposed design compares favorably to several conventional platform designs.

### Functional Concurrent Hidden Markov Model

*Xiaoxiao Zhou and ♦Xinyuan Song*

The Chinese University of Hong Kong  
xysong@sta.cuhk.edu.hk

This study considers a functional concurrent hidden Markov model. The proposed model consists of two components. One is a transition model for elucidating how potential covariates influence the transition probability from one state to another. The other is a conditional functional linear concurrent regression model for characterizing the state-specific effects of functional covariates. A distribution-free random effect is introduced to the conditional model to describe the dependency of individual functional observations. The soft-thresholding operator and the adaptive group lasso are introduced to simultaneously accommodate the local and global sparsity of the functional coefficients. A Bayesian approach is developed to jointly conduct estimation, variable selection, and the detection of zero-effect regions. This proposed approach incorporates the dependent Dirichlet process with stick-breaking prior for accommodating the

unspecified distribution of the random effect and a blocked Gibbs sampler for efficient posterior sampling. Finally, the empirical performance of the proposed method is evaluated through simulation studies, and the utility of the methodology is demonstrated by an application to the analysis of air pollution and meteorological data.

### Deep Learning for Regression Analysis of Interval-Censored Data

*Mingyue Du<sup>1</sup>, Qiang Wu<sup>2</sup>, Xingwei Tong<sup>2</sup> and ♦Xingqiu Zhao<sup>1</sup>*

<sup>1</sup>The Hong Kong Polytechnic University

<sup>2</sup>Beijing Normal University

xingqiu.zhao@polyu.edu.hk

This paper discusses regression analysis of interval-censored failure time data, and a new deep learning approach is proposed under the partially linear Cox model. For the analysis, we need to overcome theoretical and computational challenges arising from complex data structure where the partial likelihood function is no longer available. We propose to use a deep neural network and a B-spline function for approximating the nonlinear component and the baseline cumulative hazard function in the model, respectively. The proposed approach is flexible and able to circumvent the curse of dimensionality. At the same time, it facilitates the interpretability of covariate effects. The asymptotic properties of the resulting estimators are established. In particular, the finite-dimensional estimator of covariate effects is asymptotically normal and attains the semiparametric efficiency, while the deep nonparametric estimator achieves the minimax optimal rate of convergence. A simulation study is conducted to assess the finite-sample performance of the proposed approach and indicates that it works well in practical situations. Finally, the proposed method is applied to a set of real data that motivated this study.

### Session 23CHI169: Statistical Methods and Analysis for the Digital Asset Economy

#### Price Divergence in Bitcoin Market

*Dehua Shen*

Nankai University

dhs@nankai.edu.cn

This paper investigates the law of one price, price divergence, and arbitrage in financial markets by examining the determinants of inter-exchange price differentials for the Bitcoin (BTC) market. Using a unique data set for 164 major BTC exchanges that cover a period from January 2015 to May 2022, we find that country segmentation, cultural difference in inequality, dependence on others, emotional gender roles, and exchange-specific differences in systematic risk, idiosyncratic risk, return momentum and lottery-like style contribute to explain the BTC price difference between exchanges. Moreover, the blockchain factors, investor adoption, and exchange (off-chain) trading activities are significantly related to the overall price divergence in the BTC market. Additionally, the effect of national cultural differences on BTC price difference between exchanges is weaker for crypto-to-BTC pairs, but country segmentation still exists. The price divergence in the BTC market exhibit periods of large and recurrent arbitrage opportunities across exchanges.

#### Do Clean and Dirty Cryptocurrencies Connect with Financial Assets Differently? the Role of Economic Policy Uncertainty

*Kun Duan<sup>1</sup>, ♦Yingying Huang<sup>2</sup>, Yanqi Zhao<sup>1</sup> and Andrew Urquhart<sup>3</sup>*

<sup>1</sup>Huazhong University of Science and Technology

<sup>2</sup>Harbin Institute of Technology

<sup>3</sup>University of Reading

huangyingying@hit.edu.cn

This paper analyses time-varying networks of clean and dirty cryptocurrencies with green and traditional assets through a dynamic connectedness approach established by the time-varying parameter vector autoregressive (TVP-VAR) model. The underlying asymmetry of the dynamic pairwise connectedness when facing uncertainty shocks is further studied through a non-parametric quantile causality method. Our results demonstrate a limited information transmission of volatility from cryptocurrencies to both traditional and green assets, while the connection of clean cryptocurrencies (CI) with the financial system is even weaker compared to that of dirty cryptocurrencies (DI), especially after the COVID-19 pandemic. In contrast, connection within the financial system is found to be relatively closer. Moreover, causal relationships between economic policy uncertainty (EPU) and cryptocurrency-financial asset linkages are generally enhanced after the pandemic onset, while such the causality of uncertainty with DI related asset linkages tends to be even stronger. Most of the above causalities are shown to be negligible during market depression, further implying the sheltering role of the market linkages against uncertainty.

#### An Analysis of the Return – volume Relationship in Decentralised Finance (Defi)

*♦Jeffrey Chu<sup>1</sup>, Stephen Chan<sup>2</sup> and Yuanyuan Zhang<sup>3</sup>*

<sup>1</sup>Renmin University of China

<sup>2</sup>American University of Sharjah

<sup>3</sup>University of Manchester

jeffrey.jchu@ruc.edu.cn

This paper investigates the dynamic volume–return relationship of the five largest decentralised finance tokens, to better understand this relationship given the similarities with cryptocurrencies and the possible benefits for traders and practitioners. We implement the quantile-on-quantile regression and an extreme value theory approach to examine the relationship between the daily returns of the prices and trading volumes of decentralised finance tokens at varying quantiles and at the extreme tails. Our results suggest that when trading volume is experiencing large increases, the returns of the prices of tokens appear to be significantly positive for some cases but negative for others. The extreme volume-return dependence is found to be asymmetric in the extreme negative and positive tails of the distributions, where the dependence below extreme negative thresholds is essentially non-existent but above extreme positive thresholds it is significant. This asymmetric extreme dependence between returns and volume may be beneficial for developing trading strategies that incorporate trading volume data, and may indicate an inefficient market.

#### The Effect of Central Bank Digital Currency Volatility on Supply Chain Management

*Shusheng Ding*

dingshusheng@nbu.edu.cn

A Central Bank Digital Currency (CBDC) launched by the Bank of England could enable businesses to directly make electronic payments. It can be argued that digital payment is helpful in supply chain management applications. However, the adoption of CBDC in the supply chain could bring new turbulence since the CBDC value may fluctuate. Therefore, this paper intends to optimize the production plan of manufacturing supply chain based on a volatility clustering model by reducing CBDC value uncertainty. We apply both GARCH model and machine learning model to depict the CBDC volatility clustering. Empirically, we employed Baltic Dry

## Abstracts

---

Index, Bitcoin and exchange rate as main variables with sample period from 2015 to 2021 to evaluate the performance of the two models. On this basis, we reveal that our machine learning model overwhelmingly outperforms the GARCH model. Consequently, our result implies that manufacturing companies' performance can be strengthened through CBDC uncertainty reduction.

### Session 23CHI47: Advanced Statistical Methods in Omics Data Analysis

#### Model-Based Spatial Reconstruction of Large-Scale Biomolecules via Bayesian Inference of a Hierarchical Spatial Model

*Chong Shen<sup>1</sup>, Shiyu Wang<sup>2</sup>, Zhaozhi Qin<sup>2</sup> and ♦Ke Deng<sup>1</sup>*

<sup>1</sup>Tsinghua University

<sup>2</sup>Emory University  
kdeng@tsinghua.edu.cn

Revealing the spatial organization of biomolecules and characterizing their spatial distribution in cells and tissues have long been recognized as important problems in biomedical research. With rapid advances of DNA sequencing technologies in recent years, creative sequencing-based experimental assays, e.g., Hi-C and DNA microscopy, have been invented to reveal the spatial properties of large-scale biomolecules in a high-throughput and high-resolution manner. A typical experiment based on these technologies produces a count matrix to record the contact frequencies among molecules of interest, which are closely associated to their spatial distances, allowing us to reconstruct the spatial organization of large-scale biomolecules via data analysis. There is a great appeal to develop statistically rigorous and computationally scalable methods for this important problem. In this study, we fill in this gap with a novel method named HiSpa. Equipped with a hierarchical spatial model, HiSpa utilizes the idea of multi-scale modelling to reduce the computation complexity from  $O(n^2)$  to  $O(n^{3/2})$  with little loss on the quality of the reconstructed spatial structure. Advanced Monte Carlo strategies are developed for efficient Bayesian inference of HiSpa. Superiority of HiSpa over existing methods is demonstrated by simulation studies and real data applications.

#### Scemail: Universal and Source-Free Annotation Method for ScRNA-Seq Data with Novel Cell-Type Perception

*Hui Wan<sup>1</sup>, Liang Chen<sup>2</sup> and ♦Minghua Deng<sup>1</sup>*

<sup>1</sup>Peking University

<sup>2</sup>Huawei Technologies Co.  
dengmh@math.pku.edu.cn

Current cell-type annotation tools for single-cell RNA sequencing (scRNA-seq) data mainly utilize well-annotated source data to help identify cell types in target data. However, on account of privacy preservation, their requirements for raw source data may not always be satisfied. In this case, achieving feature alignment between source and target data explicitly is impossible. Additionally, these methods are barely able to discover the presence of novel cell types. A subjective threshold is often selected by users to detect novel cells. We propose a universal annotation framework for scRNA-seq data called scEMAIL, which automatically detects novel cell types without accessing source data during adaptation. For new cell-type identification, a novel cell-type perception module is designed. Model adaptation is then conducted to alleviate the batch effect. We gather multi-order neighborhood messages globally and impose local affinity regularizations on "known" cells. These constraints mitigate wrong classifications of the source model via reli-

able self-supervised information of neighbors. scEMAIL is accurate and robust under various scenarios in both simulation and real data. It is also flexible to be applied to challenging single-cell ATAC-seq data without loss of superiority.

#### Integrative Analysis of 16S Marker-Gene and Shotgun Metagenomic Sequencing Data Improves the Efficiency of Testing Hypotheses About the Microbiome

*Ye Yue, Glen Satten and ♦Yijuan Hu*

Emory University  
yijuan.hu@emory.edu

The most widely used technologies for profiling microbial communities are 16S marker-gene sequencing and shotgun metagenomic sequencing (SMS). Each experiment is subject to unique, systematic experimental biases that are introduced in every step of the experiment, which result in systematically distorted observations of the true underlying microbial composition within a sample. Many microbiome studies have both 16S and SMS data on the same cohort of samples, both of which provide consistent patterns in taxonomic profiles and can be integrated for improved power of testing the association of these patterns with sample-level covariates. In this article, we present the first method for such integrative analysis of 16S and SMS data. The method is based on our LOCOM model (Hu et al., 2022, PNAS), which uses logistic regression for testing differential abundance of taxa that is robust to experimental bias. We extend LOCOM to allow differential experimental bias and assign data-adaptive weights to observations from the two experiments. We demonstrate the superiority of the new method over alternative approaches via extensive simulation results and application to two real studies found in Qiita.

### Session 23CHI48: Emerging Development in Statistical Analyses for Multi-Omic Data

#### Bayesian Integrative Region Segmentation in Spatially Resolved Transcriptomic Studies

*♦Yinqiao Yan and Xiangyu Luo*

Renmin University of China  
yanyin@ruc.edu.cn

The spatially resolved transcriptomic study is a recently developed biological experiment that can measure gene expressions and retain spatial information simultaneously, opening a new avenue to characterize fine-grained tissue structures. In this article, we propose a nonparametric Bayesian method named BINRES to carry out the region segmentation for a tissue section by integrating all the three types of data generated during the study—gene expressions, spatial coordinates, and the histology image. BINRES is able to capture more subtle regions than existing statistical partitioning models that only partially make use of the three data modes and is more interpretable than neural-network-based region segmentation approaches. Specifically, due to a nonparametric spatial prior, BINRES does not require a prespecified region number and can learn it automatically. BINRES also combines the image and the gene expressions in the Bayesian consensus clustering framework and thus flexibly adjusts their contribution weights in a data-adaptive manner. A computationally scalable extension is developed for large-scale studies. Both simulation studies and the real application to three mouse spatial transcriptomic datasets demonstrate that BINRES outperforms the competing methods and easily achieves the uncertainty quantification of the integrative partition.

#### Enhancing the Study of Microbiome-Metabolome Interactions:

## a Transfer-Learning Approach for Precise Identification of Essential Microbes

♦Yue Wang<sup>1</sup>, Lillian Li<sup>2</sup>, Chenglong Ye<sup>2</sup> and Tim Randolph<sup>3</sup>

<sup>1</sup>Colorado School of Public Health

<sup>2</sup>University of Kentucky

<sup>3</sup>Fred Hutch Cancer Center

yue.2.wang@cuanschutz.edu

Recent research has revealed the essential role that microbial metabolites play in host-microbiome interactions. Although statistical and machine-learning methods have been employed to explore microbiome-metabolome interactions in multiview microbiome studies, most of these approaches focus solely on the prediction of microbial metabolites, which lacks biological interpretation. Additionally, existing methods face limitations in either prediction or inference due to small sample sizes and highly correlated microbes and metabolites. To overcome these limitations, we present a transfer-learning method that evaluates microbiome-metabolome interactions. Our approach efficiently utilizes information from comparable metabolites obtained through external databases or data-driven methods, resulting in more precise predictions of microbial metabolites and identification of essential microbes involved in each microbial metabolite. Our numerical studies demonstrate that our method enables a deeper understanding of the mechanism of host-microbiome interactions and establishes a statistical basis for potential microbiome-based therapies for various human diseases.

## A Novel Transcriptome-Wide Association Study Method with Incorporating Multiple Annotations

Han Wang and ♦Yan Dora Zhang

The University of Hong Kong  
doraz@hku.hk

Transcriptome-wide association studies (TWAS) aims to identify regulated genes that affect diseases, is critical in understanding the molecular mechanism of complex traits or human diseases. Classic TWAS involve two steps: first conduct a linear regression model based on a gene expression panel, and then conduct gene-trait association test. In this talk, motivated by the fact that SNPs with actively expressed epigenetic annotations may be more likely to regulate gene expressions, we propose to incorporate annotation information into the first step of TWAS model through a nonparametric Bayesian prior to improve the accuracy of imputation model. The proposed new method leads to a better predictive performance and higher power in simulations. Real analysis results based on several summary GWAS data will also be introduced.

## Me-Bayes SL: Enhanced Bayesian Polygenic Risk Prediction Leveraging Information Across Multiple Ancestry Groups

♦Jin Jin<sup>1</sup>, Jianan Zhan<sup>2</sup>, Jingning Zhang<sup>3</sup>, Ruzhang Zhao<sup>3</sup>, Jared O'connell<sup>2</sup>, Yunxuan Jiang<sup>2</sup>, Steven Buyske<sup>4</sup>, Genevieve Wojcik<sup>5</sup>, Haoyu Zhang<sup>6</sup> and Nilanjan Chatterjee<sup>7</sup>

<sup>1</sup>Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health; Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania

<sup>2</sup>23andMe, Inc.

<sup>3</sup>Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

<sup>4</sup>Department of Statistics, Rutgers University

<sup>5</sup>Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health

<sup>6</sup>Division of Cancer Epidemiology and Genetics, National Cancer Institute

<sup>7</sup>Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health; Department of Oncology, School of Medicine, Johns Hopkins University  
jin.jin@pennmedicine.upenn.edu

Polygenic risk scores (PRS) are now showing promising predictive performance on a wide variety of complex traits and diseases, but there exists a substantial performance gap across different populations. We propose ME-Bayes SL, a novel method for ancestry-specific polygenic risk prediction that borrows information in the summary statistics from genome-wide association studies (GWAS) across multiple ancestry groups. ME-Bayes SL conducts Bayesian hierarchical modeling under a multivariate spike-and-slab model for effect-size distribution and incorporates an ensemble learning step to combine information across different tuning parameter settings and ancestry groups. In our data analyses of 16 traits across four distinct studies, totaling 5.7 million participants with a substantial ancestral diversity, ME-Bayes SL shows promising performance compared to alternatives. The method, for example, has an average gain in prediction R2 across 11 continuous traits of 40.2% and 49.3% compared to PRS-CSx and CT-SLEB, respectively, in the African Ancestry population. The best-performing method, however, varies by GWAS sample size, target ancestry group, underlying trait architecture, and the choice of reference samples for LD estimation, and thus ultimately, a combination of methods may be needed to generate the most robust PRS across diverse populations.

## Session 23CHI50: Recent Advances in Data Fusion and Integration with Real-World Applications in Healthcare

### Robust Machine Learner for Mean Potential Outcome with Information Integration from Auxiliary Data

♦Chixiang Chen<sup>1</sup>, Shuo Chen, Zhenyao Ye, Xu Shi and Tianzhou Ma

<sup>1</sup>university of maryland, school of medicine  
chixiang.chen@som.umaryland.edu

Modern machine learning algorithms show powerful performance in prediction problems. However, owing to their black-box nature, it is difficult to statistically evaluate their performance, which could substantially vary across databases and underlying set-ups. As a result, locating the most appropriate algorithm may be very challenging when multiple algorithms are considered and evaluated, especially in the context of causal inference. In this talk, the first goal is to describe a new and robust machine learner for mean potential outcome. The learner enables valid statistical inference and has the property of robustness: multiple machine learning algorithms can be incorporated, and robustness is retained as long as one of the candidate algorithms works well for propensity score and one works well for conditional mean, without requiring an independent test data. The second goal is to illustrate a new data-integration scheme for synthesizing extra information from auxiliary data into the machine learner to substantially boost the estimation efficiency. Extensive numerical studies and a real data application of the impact of substance use on brain ageing demonstrate the superiority of our method over competing methods.

### Lossless One-Shot Distributed Linear Mixed Model for a Multi-Site International Study of Covid-19 Hospitalization Length of Stay

♦Chongliang Luo<sup>1</sup>, Md. Nazmul Islam<sup>2</sup>, Jenna Reps<sup>3</sup>, Rui Duan<sup>4</sup>, Jiang Bian<sup>5</sup>, Talita Duarte-Salles<sup>6</sup>, Thomas Falconer<sup>7</sup>, Chungsoo Kim<sup>8</sup>, Hua Xu<sup>9</sup> and Yong Chen<sup>10</sup>

## Abstracts

---

<sup>1</sup>Washington University in St Louis

<sup>2</sup>Optum Labs

<sup>3</sup>Janssen Research and Development LLC

<sup>4</sup>Harvard T.H. Chan School of Public Health,

<sup>5</sup>University of Florida

<sup>6</sup>Fundacio Institut Universitari per a la recerca a l'Atencio Primaria de Salut Jordi Gol i Gurina (IDIAPJGol)

<sup>7</sup>Columbia University

<sup>8</sup>Ajou University Graduate School of Medicine

<sup>9</sup>The University of Texas Health Science Center at Houston

<sup>10</sup>University of Pennsylvania  
chongliang@wustl.edu

Linear mixed models are commonly used in healthcare-based association analyses for analyzing multi-site data with heterogeneous site-specific random effects. Due to regulations for protecting patients' privacy, sensitive individual patient data (IPD) typically cannot be shared across sites. We propose an algorithm for fitting distributed linear mixed models (DLMMs) without sharing IPD across sites. This algorithm achieves results identical to those achieved using pooled IPD from multiple sites (i.e., the same effect size and standard error estimates), hence demonstrating the lossless property. The algorithm requires each site to contribute minimal aggregated data in only one round of communication. We demonstrate the lossless property of the proposed DLMM algorithm by investigating the associations between demographic and clinical characteristics and length of hospital stay in COVID-19 patients using administrative claims from the UnitedHealth Group Clinical Discovery Database. We extend this association study by incorporating 120,609 COVID-19 patients from 11 collaborative data sources worldwide.

### Tree-Guided Rare Feature Selection and Logic Aggregation with Electronic Health Records Data

Kun Chen

University of Connecticut  
kun.chen@uconn.edu

Statistical learning with a large number of rare binary features is commonly encountered in analyzing electronic health records (EHR) data. Dealing with the resulting highly sparse and large-scale binary feature matrix is notoriously challenging as conventional methods may suffer from a lack of power in testing and inconsistency in model fitting, while machine learning methods may suffer from the inability to produce interpretable results. To improve EHR-based modeling and utilize the natural hierarchical structure of disease classification, we propose a tree-guided feature selection and logic aggregation approach for large-scale regression with rare binary features, in which dimension reduction is achieved through not only a sparsity pursuit but also an aggregation promoter with the logic operator of "or". We convert the combinatorial problem into a convex linearly-constrained regularized estimation, which enables scalable computation with theoretical guarantees. In a suicide risk study with EHR data, our approach is able to select and aggregate prior mental health diagnoses as guided by the diagnosis hierarchy of the International Classification of Diseases. By balancing the rarity and specificity of the EHR diagnosis records, our strategy improves both prediction and interpretation.

### Session 23CHI51: Big Data Analysis Based on Statistics and Machine Learning

#### Statistical Analysis of Civil Aviation Big Data

Riquan Zhang

School of Statistics and Information, Shanghai University of International Business and Economics  
zhangriquan@163.com

Aviation safety big data mainly includes flight data, cloud license data, meteorological data, geographic data, and safety information data, and has characteristics such as multi-source, massive, heterogeneous, low-quality, missing, time-varying, real-time, and dependent. It is imperative to use big data statistical analysis to achieve intelligent and accurate data driven aviation safety management. This report introduces aviation safety big data and some key scientific issues it faces based on, and some achievements in flight quality monitoring, quantitative analysis of typical event risks, intelligent early warning methods, and evaluation of pilot flight control capabilities were studied. In response to the challenges and difficulties currently faced in aviation safety big data, some research directions were proposed from four aspects: multi-source data matching and fusion, machine theory and methods of popular structures, attribution and traceability of safety events, and real-time flight data analysis.

#### Forecasting Stock Volatility and Value-at-Risk Based on Temporal Convolutional Networks

♦Chunxia Zhang<sup>1</sup>, Jun Li<sup>1</sup>, Xingshang Huang<sup>2</sup>, Jiangshe Zhang<sup>1</sup> and Huachuan Huang<sup>1</sup>

<sup>1</sup>Xi'an Jiaotong University

<sup>2</sup>Nanjing Audit University  
30989108@qq.com

Considering that the volatility of financial asset returns as well as the Value-at-Risk (VaR) play a significant role in many applications such as risk management, investment portfolios and etc. , it is thus extremely essential to accurately estimate volatility and VaR. In this talk, we will introduce how to utilize Temporal convolutional networks (TCNs) to forecast stock volatility and VaR. In the experiments conducted with both synthetic data and some real stock data, TCNs are compared with other twelve popular models which include nine conventional approaches (i.e., three GARCH-type models with each being considered three tail distributions) and three deep learning methods (i.e., LSTM, LSTM with attention mechanism and GRU). The Friedman test followed by the Nemenyi post-hoc test is also employed to analyze whether TCNs perform significantly better than the other methods across the real stock datasets. As for volatility modeling, experimental results show that TCNs outperforms all the other methods in terms of RMSE and MAE. In the meantime, TCNs behave best in calculating VaR when evaluating their performance with several metrics. As a result, TCNs can be regarded as an important technique to forecast return volatility and the associated VaR.

#### Deep Convolutional Neural Network with Feature Ensemble for Image Classification

Yu Wang

Shanxi University  
wangyu@sxu.edu.cn

Deep convolutional neural networks such as AlexNet, VGGNet and ResNet have been extensively employed in image classification task. A common solution is directly feeding deep CNN (Convolutional Neural Network) features extracted from a deep network into a classification function. However, this solution may easily result in poor accuracy and robustness due to the single experimental result. One alternative is utilizing ensemble of multiple deep networks. And this would bring very expensive, even unacceptable computational complexity. Thus, we propose a new deep CNN feature ensemble frame based on multiple cross validation resampling results of the single feature layer to cope with the above two

issues. Theoretically, the proposed method is proved that having a smaller error rate and a lower Rademacher complexity than the single feature layer method. Moreover, extensive experiments on several challenging image classification data sets demonstrate the superiority of the proposed method.

### An Improved SNE with Its Applications in Classification and Visualization

*Peilin Sun and ♦Xu Qin*

University of Electronic Science and Technology of China  
qinxu@uestc.edu.cn

Data embedding aims at maintaining the complete information of original data so that the difference between the original and the embedded data is imperceptible. Stochastic Neighbor Embedding(SNE) as a nonlinear manifold learning algorithm has received extensive attention. Considering the multimodality of actual data, we propose an improved Stochastic Neighbor Embedding, SL3-SNE. The technique is a variation of Stochastic Neighbor Embedding that produces better clusterings by introducing spherical logistic distribution, which is more heavy-tailed than normal distribution and is able to characterize the multimodality of data. We demonstrate the performance of SL3-SNE on a variety of simulated data and real data and compare it with t-SNE and vMF-SNE. The visualizations and classification of SL3-SNE are significantly better than others because it not only clusters well, but also does a good job in classificaion prediction tasks.

### Session 23CHI55: New Advancements in Analytical Methods using Multi-Source or Multi-Trial Data

#### Multiply-Robust Estimation of Causal Treatment Effect on a Binary Outcome with Integrated Information from Secondary Outcomes

*Chixiang Chen<sup>1</sup>, Shuo Chen<sup>1</sup>, Qi Long<sup>2</sup>, Sudeshna Das<sup>3</sup> and ♦Ming Wang<sup>4</sup>*

<sup>1</sup>University of Maryland

<sup>2</sup>University of Pennsylvania

<sup>3</sup>Harvard Medical School

<sup>4</sup>Case Western Reserve University  
mxw827@case.edu

An assessment of the causal treatment effect in the development and progression of certain diseases is important in clinical trials and biomedical studies. However, it is not possible to infer a causal relationship when the treatment assignment is imbalanced and confounded by other mechanisms. Specifically, when the treatment assignment is not randomized and the primary outcome is binary, a conventional logistic regression may not be valid to elucidate any causal inference. Moreover, exclusively capturing all confounders is extremely difficult and even impossible in large-scale observational studies. We propose a multiply-robust (MultiR) estimator for estimating the causal effect with a binary outcome, where multiple propensity score models and conditional mean imputation models are used to ensure estimation robustness. Further, we propose an enhanced MultiR (eMultiR) estimator that reduces the estimation variability of MultiR estimates by incorporating secondary outcomes that are highly correlated with the primary binary outcome. The resulting estimates are less sensitive to model mis-specification compared to those based on state-of-the-art methods (e.g., doubly-robust estimators). These estimates are verified through both theoretical and numerical assessments. Finally, the utility of (e)MultiR

estimation is illustrated using the Uniform Data Set (UDS) from the National Alzheimer's Coordinating Center.

#### Interpretable Image Segmentation Network Originated from Multi-Grid Variational Model

*Junying Meng<sup>1</sup>, ♦Weihong Guo<sup>2</sup>, Jun Liu<sup>1</sup> and Mingrui Yang<sup>3</sup>*

<sup>1</sup>Beijing Normal University

<sup>2</sup>Case Western Reserve University

<sup>3</sup>Cleveland Clinic  
wxg49@case.edu

Image segmentation divides an image domain into homogeneous regions for further analysis. This is a significant and crucial task in various applications such as medical imaging. Deep learning (DL) methods have been proposed and widely used for image segmentation. However, these methods require a large amount of manually segmented data as training data and suffer from poor interpretability. The classical Mumford-Shah (MS) model is effective for segmentation and provides a piece-wise smooth approximation of the original image. In this paper, we replace the hand-crafted regularity term in the MS model with a data adaptive generalized learnable regularity term and use a multi-grid framework to unroll the MS model and obtain a variational model-based segmentation network with better generalizability and interpretability. This approach allows for the incorporation of learnable prior information into the network structure design. Moreover, the multi-grid framework enables multi-scale feature extraction and offers a mathematical explanation for the effectiveness of the U-shaped network structure in producing good image segmentation results. Since the proposed network originates from a variational model, it can also handle small training sizes. Our experiments on various data sets testify the advantages over some other methods.

#### Statistical Concerns in Meta-Analysis of Safety Data

*Shouhao Zhou*

Penn State University

szhou1@pennstatehealth.psu.edu

Information on drug safety is critical to health care and policy decisions, as treatment recommendations hinge on accurate knowledge of both efficacy and harms. However, assessments of drug safety from individual clinical trials are often underpowered due to insufficient sample sizes and have limited generalizability due to restrictive inclusion/exclusion criteria. Meta-analysis offers a unique opportunity to assess adverse event risks in a large sample size for a broad population, but poses several challenges and requires careful consideration. In particular, many adverse events are rare, and adverse events which occur infrequently may be omitted from official trial reports. The observed data therefore correspond to low counts which may be subject to informative censoring. We propose to develop a novel Bayesian modeling framework to synthesize toxicity data. It can handle hierarchical multivariate outcome data and allow for incomplete reporting. We illustrate our method by applying it to a meta-analysis of adverse events in cancer immunotherapy and targeted therapy.

#### Semiparametric and Variable Marginal Effects with Debiased Machine Learning

*♦Chan Shen<sup>1</sup> and Roger Klein<sup>2</sup>*

<sup>1</sup>Penn State University

<sup>2</sup>Rutgers University

cshen@pennstatehealth.psu.edu

Machine learning used to be mainly focused on prediction as the methods are often vulnerable to biases when we try to apply it to causal inference. There have been some recent developments

## Abstracts

---

on using machine learning to recover marginal effects (e.g. Chernozhukov et. al.). Such methods usually involve debiasing mechanisms and assume that the marginal effects are constant. In this study, we propose a semiparametric approach with debiased machine learning to estimate marginal effects that can depend on covariates through unknown functions of indices. Our Monte Carlo study shows desirable properties of this semiparametric approach.

### Session 23CHI57: Emerging Problems and Solutions in Imaging Statistics

#### Survival Models with Medical Images as Predictors

Zhangsheng Yu

Shanghai Jiao Tong University  
yuzhangsheng@sjtu.edu.cn

I will first introduce the deep survival model application to predict the survival of stomach cancer patients using pathology images, genetics data, and clinical data. I will also introduce a partial linear Cox model for interval-censored data with deep neural networks where the nonparametric component is estimated with a deep ReLU network. This model not only retains the nice interpretability of the parametric component but also improves the prediction power compared with the partial linear additive Cox model. We derive the convergence rate of the proposed estimator and show that it can break curse of dimensionality under some smoothness assumptions. Base on such rate, the asymptotic normality and the semiparametric efficiency are also established. Intensive simulation studies are carried out to demonstrate the finite sample performance on both estimation and prediction.

#### Latent Subgroup Identification in Image-on-Scalar Regression

Zikai Lin, ♦Yajuan Si and Jian Kang

University of Michigan  
yajuan@umich.edu

Image-on-scalar regression has been a popular approach to modeling the association between brain activities and scalar characteristics in neuroimaging research. The associations could be heterogeneous across individuals in the population, as indicated by recent large-scale neuroimaging studies, e.g., the Adolescent Brain Cognitive Development (ABCD) study. The ABCD data can inform our understanding of heterogeneous associations and how to leverage the heterogeneity and tailor interventions to increase the number of youth who benefit. It is of great interest to identify subgroups of individuals from the population such that: 1) within each subgroup the brain activities have homogeneous associations with the clinical measures; 2) across subgroups the associations are heterogeneous; and 3) the group allocation depends on individual characteristics. Existing image-on-scalar regression methods and clustering methods cannot directly achieve this goal. We propose a latent subgroup image-on-scalar regression model (LASIR) to analyze large-scale, multi-site neuroimaging data with diverse sociodemographics. LASIR introduces the latent subgroup for each individual and group-specific, spatially varying effects, with an efficient stochastic expectation maximization algorithm for inferences. We demonstrate that LASIR outperforms existing alternatives for subgroup identification of brain activation patterns with functional magnetic resonance imaging data via comprehensive simulations and applications to the ABCD study.

#### Mediation Analysis for High-Dimensional Mediators and Outcomes with an Application to Multimodal Imaging Data

Zhiwei Zhao and ♦Shuo Chen

University of Maryland  
shuochen@som.umd.edu

Multimodal neuroimaging data have attracted increasing attention for brain research. An integrated analysis of multimodal neuroimaging data and behavioral or clinical measurements provides a promising approach for comprehensively and systematically investigating the underlying neural mechanisms of different phenotypes. However, such an integrated data analysis is intrinsically challenging due to the complex interactive relationships between the multimodal multivariate imaging variables. To address this challenge, a novel multivariate-mediator and multivariate-outcome mediation model (MMO) is proposed to simultaneously extract the latent systematic mediation patterns and estimate the mediation effects based on a dense bi-cluster graph approach. A computationally efficient algorithm is developed for dense bi-cluster structure estimation and inference to identify the mediation patterns with multiple testing correction. The performance of the proposed method is evaluated by an extensive simulation analysis with comparison to the existing methods. The results show that MMO performs better in terms of both the false discovery rate and sensitivity compared to existing models. The MMO is applied to a multimodal imaging dataset from the Human Connectome Project to investigate the effect of systolic blood pressure on whole-brain imaging measures for the regional homogeneity of the blood oxygenation level-dependent signal through the cerebral blood flow.

#### Statistical Inferences for Complex Dependence of Multimodal Imaging Data

Jinyuan Chang<sup>1</sup>, ♦Jing He<sup>1</sup>, Jian Kang<sup>2</sup> and Mingcong Wu<sup>1</sup>

<sup>1</sup>Southwestern University of Finance and Economics

<sup>2</sup>University of Michigan  
he\_jing@swufe.edu.cn

Statistical analysis of multimodal imaging data is a challenging task, since the data involves high-dimensionality, strong spatial correlations and complex data structures. In this work, we propose rigorous statistical testing procedures for making inferences on the complex dependence of multimodal imaging data. Motivated by the analysis of multi-task fMRI data in the Human Connectome Project (HCP) study, we particularly address three hypothesis testing problems: (a) testing independence among imaging modalities over brain regions, (b) testing independence between brain regions within imaging modalities, and (c) testing independence between brain regions across different modalities. Considering a general form for all the three tests, we develop a global testing procedure and a multiple testing procedure controlling the false discovery rate. We study theoretical properties of the proposed tests and develop a computationally efficient distributed algorithm. The proposed methods and theory are general and relevant for many statistical problems of testing independence structure among the components of high-dimensional random vectors with arbitrary dependence structures. We also illustrate our proposed methods via extensive simulations and analysis of five task fMRI contrast maps in the HCP study.

### Session 23CHI7: Data Borrowing : Methodology and Application

#### Bayesian Borrowing from Historical Control Data in Vaccine Efficacy Trial

Penny Peng

Department of Biostatistics and Programming, China, Sanofi, Inc.  
Penny1.peng@sanofi.com

In the context of vaccine efficacy trial where the incidence rate is very low and a very large sample size is usually expected, incorporating historical data into a new trial is extremely attractive to reduce sample size and increase estimation precision. Nevertheless, for some infectious diseases, seasonal change in incidence rates poses a huge challenge in borrowing historical data and a critical question is how to properly take advantage of historical data borrowing with acceptable tolerance to between-trials heterogeneity commonly from seasonal disease transmission. In this article, we extend a probability-based power prior which determines the amount of information to be borrowed based on the agreement between the historical and current data, to make it applicable for either a single or multiple historical trials available, with constraint on the amount of historical information to be borrowed. Simulations are conducted to compare the performance of the proposed method with other methods including modified power prior (MPP), meta-analytic-predictive (MAP) prior and the commensurate prior methods. Furthermore, we illustrate the application of the proposed method for trial design in a practical setting.

#### **A Brief Introduction of Historical Data Borrowing Approach Implemented in a Real Case Study.**

♦*Gaowei Nian, Ning Li and Genming Shi*

Gaowei.Nian@Sanofi.com

Historical or external data borrowing has demonstrated its main advantages to increase the statistical power and improve the quantitative decision marking in the clinical trial and been discussed in large amounts of literature and regulatory guidelines. This presentation will briefly introduce the historical data borrowing Bayesian approach, and discuss the benefits and risks. The implementation of borrowing historical data will be demonstrated in a trial with the historical placebo data borrowing as a supplementary approach.

#### **Methods of Reconstructing Individual Patients Data and Subgroup Survival Curves from Published Kaplan-Meier Plots**

*Sheng Xu*

BeiGene

sheng.xu@beigene.com

Utilizing the published survival curves is a common occurrence in cost-effectiveness analysis, indirect treatment comparison, comparator data mining, among other quantitative analyses for commercial objectives in pharmaceutical industry. Recovery of the individual patient data (IPD) can enhance the analytical power. In the talk, we will review and compare the methods of IPD reconstruction with and without incorporating censoring marks under different scenarios. We will demonstrate some applications of the secondary data analysis using reconstructed pseudo IPD. Occasionally, a manuscript publishes only the survival curves of the full efficacy analysis set and one of the subgroups, but one is interested in the other unreported subgroup. In this case, we present a pipeline of subgroup survival curve recovery by the combination of different digitization, IPD reconstruction, and matching techniques. A real-world example will be demonstrated.

#### **Reducing Bias using Propensity Score-Integrated Composite Likelihood Approach for Incorporating Multiple External Data Sources**

♦*Leixin Xia<sup>1</sup>, Yishen Yang<sup>2</sup>, Weilong Zhao<sup>1</sup>, Chaohui Yuan<sup>1</sup>, Ming Chen<sup>1</sup>, Wei Tan<sup>3</sup>, Zhini Wang<sup>2</sup> and Bin Jia<sup>1</sup>*

<sup>1</sup>Janssen

<sup>2</sup>IQVIA

<sup>3</sup>ICON plc

lxia17@its.jnj.com

A propensity score-integrated composite likelihood approach is developed to augment the control arm of a two-arm randomized controlled trial (RCT) with subjects from multiple external data sources such as real-world data (RWD) and historical clinical studies containing subject-level outcomes and covariates. The propensity scores for the subjects in the external data sources versus the subjects in the RCT are first estimated, and then subjects are placed in different strata based on their estimated propensity scores. Within each propensity score stratum, a composite likelihood is formulated with the information contributed by the external data sources, and inference on the treatment effect is obtained. The proposed approach is implemented under the two-stage study design framework utilizing the outcome-free principle to ensure the integrity of a study. Simulation study is provided to demonstrate the implementation of the proposed approach

#### **Session 23CHI73: Recent Developments in Medical Bioinformatics**

##### **Biomedical Entity Linking by Text Generation and Knowledge Enhancement**

♦*Hongyi Yuan, Zheng Yuan and Sheng Yu*

Tsinghua University

yuanhy20@mails.tsinghua.edu.cn

Entities lie in the heart of biomedical natural language understanding, and the biomedical entity linking (EL) task remains challenging due to the fine-grained and diversiform concept names. Generative methods achieve remarkable performances in general domain EL with less memory usage while requiring expensive pre-training. However, it is not trivial to take advantage of a generative method, due to the lack of biomedical entity knowledge in a generative model. This work uses a generative approach to model biomedical EL and proposes KB-guided pre-training and continuous language model pre-training to inject entity knowledge. KB-guided pre-training is implemented by constructing synthetic samples with synonyms and definitions from KB and continuous language model pre-training uses general pre-training methods on PubMed articles to adapt to biomedical domain knowledge. We also propose synonyms-aware fine-tuning to select concept names for training and propose a decoder prompt and multi-synonyms constrained prefix tree for inference. Entity linking performance can be enhanced with both entity knowledge injecting methods, and achieve state-of-the-art linking performance.

##### **Multimodal Learning on Graphs for Disease Relation Extraction**

♦*Yucong Lin<sup>1</sup>, Keming Lu<sup>2</sup>, Sheng Yu<sup>3</sup>, Tianxi Cai<sup>4</sup> and Marinka Zitnik<sup>4</sup>*

<sup>1</sup>aInstitute of Engineering Medicine, Beijing Institute of Technology, Beijing, China

<sup>2</sup>cViterbi School of Engineering, University of Southern California, Los Angeles, USA

<sup>3</sup>eDepartment of Industrial Engineering, Tsinghua University, Beijing, China

<sup>4</sup>gDepartment of Biomedical Informatics, Harvard Medical School, Boston, MA, 02115, USA  
40796884@qq.com

Objective: Disease knowledge graphs have emerged as a powerful tool for AI, enabling the connection, organization, and access to diverse information about diseases. However, the relations between disease concepts are often distributed across multiple data

## Abstracts

formats, including plain language and incomplete disease knowledge graphs. As a result, extracting disease relations from multimodal data sources is crucial for constructing accurate and comprehensive disease knowledge graphs. Methods: We introduce REMAP, a multimodal approach for disease relation extraction. The REMAP machine learning approach jointly embeds a partial, incomplete knowledge graph and a medical language dataset into a compact latent vector space, aligning the multimodal embeddings for optimal disease relation extraction. Additionally, REMAP utilizes a decoupled model structure to enable inference in single-modal data, which can be applied under missing modality scenarios. Results: We apply the REMAP approach to a disease knowledge graph with 96,913 relations and a text dataset of 1.24 million sentences. On a dataset annotated by human experts, REMAP improves language-based disease relation extraction by 10.0% (accuracy) and 17.2% (F1-score) by fusing disease knowledge graphs with language information. Furthermore, REMAP leverages text information to recommend new relationships in the knowledge graph, outperforming graph-based methods by 8.4% (accuracy) and 10.4% (F1-score).

### Knowledge Graph Embedding with Electronic Health Records Data

♦*Junwei Lu, Tianxi Cai and Doudou Zhou*

Harvard T.H. Chan School of Public Health  
junweilu@hsph.harvard.edu

Due to the increasing adoption of electronic health records (EHR), large scale EHRs have become another rich data source for translational clinical research. We propose to infer the conditional dependency structure among EHR features via a latent graphical block model (LGBM). The LGBM has a two layer structure with the first providing semantic embedding vector (SEV) representation for the EHR features and the second overlaying a graphical block model on the latent SEVs. The block structures on the graphical model also allow us to cluster synonymous features in EHR. We propose to learn the LGBM efficiently, in both statistical and computational sense, based on the empirical point mutual information matrix. We establish the statistical rates of the proposed estimators and show the perfect recovery of the block structure. Numerical results from simulation studies and real EHR data analyses suggest that the proposed LGBM estimator performs well in finite sample.

### RWE-Ready: Pipeline to Harness Electronic Health Records for Real-World Evidence

♦*Jue Hou<sup>1</sup>, Rachel Zhao<sup>2</sup>, Jessica Gronsbell<sup>3</sup>, Yucong Lin<sup>4</sup>, Clara-Lea Bonzel<sup>5</sup>, Qingyi Zeng<sup>1</sup>, Sinian Zhang<sup>6</sup>, Brett Beaulieu-Jones<sup>7</sup>, Griffin Webber<sup>5</sup> and And Others<sup>8</sup>*

<sup>1</sup>University of Minnesota

<sup>2</sup>University of British Columbia

<sup>3</sup>University of Toronto

<sup>4</sup>Beijing Institute of Technology

<sup>5</sup>Harvard Medical School

<sup>6</sup>Renmin University of China

<sup>7</sup>University of Chicago

<sup>8</sup>Merck & Co, Duke University, Harvard Medical School  
hou00123@umn.edu

While randomized controlled trials (RCTs) are the gold-standard for establishing the efficacy and safety of a medical treatment, real-world evidence (RWE) generated from real-world data (RWD) has been vital in post-approval monitoring and is being promoted for the regulatory process of experimental therapies. An emerging source of RWD is electronic health records (EHRs), which contain de-

tailed information on patient care in both structured (e. g., diagnosis codes) and unstructured (e. g., clinical notes, images) form. Despite the granularity of the data available in EHRs, critical variables required to reliably assess the relationship between a treatment and clinical outcome can be challenging to extract. We provide an integrated data curation and modeling pipeline with four modules leveraging recent advances in natural language processing (NLP), computational phenotyping, causal modeling techniques with noisy data to address this fundamental challenge and accelerate the reliable use of EHRs for RWE.

### Session 23CHI77: Recent Advances in Biostatistics

#### Sc-Meb: Spatial Clustering with Hidden Markov Random Field using Empirical Bayes

♦*Yang Yi<sup>1</sup>, Xingjie Shi<sup>2</sup>, Wei Liu<sup>1</sup>, Qiuzhong Zhou<sup>3</sup>, Mai Chan Lau<sup>4</sup>, Jeffrey Chun Tatt Lim<sup>4</sup>, Lei Sun<sup>5</sup>, Cedric Chuan Young Ng<sup>6</sup>, Joe Yeong<sup>7</sup> and Liu Liu<sup>8</sup>*

<sup>1</sup>Health Services & Systems Research program at Duke-NUS Medical School

<sup>2</sup>Academy of Statistics and Interdisciplinary Sciences at East China Normal University, Shanghai, China

<sup>3</sup>Cardiovascular and Metabolic Disorders program at Duke-NUS Medical School.

<sup>4</sup>IMCB, ASTAR Singapore.

<sup>5</sup>Cardiovascular and Metabolic Disorders program at Duke-NUS Medical School, Singapore.

<sup>6</sup>CTO of Cancer Discovery Hub, National Cancer Centre Singapore

<sup>7</sup>IMCB, ASTAR Singapore

<sup>8</sup>Health Services & Systems Research program at Duke-NUS Medical School, Singapore.  
yangyi2000603@gmail.com

Spatial transcriptomics has been emerging as a powerful technique for resolving gene expression profiles while retaining tissue spatial information. These spatially resolved transcriptomics make it feasible to examine the complex multicellular systems of different microenvironments. To answer scientific questions with spatial transcriptomics and expand our understanding of how cell types and states are regulated by microenvironment, the first step is to identify cell clusters by integrating the available spatial information. Here, we introduce SC-MEB, an empirical Bayes approach for spatial clustering analysis using a hidden Markov random field. We have also derived an efficient expectation-maximization algorithm based on an iterative conditional mode for SC-MEB. In contrast to BayesSpace, a recently developed method, SC-MEB is not only computationally efficient and scalable to large sample sizes but is also capable of choosing the smoothness parameter and the number of clusters. We performed comprehensive simulation studies to demonstrate the superiority of SC-MEB over some existing methods. We applied SC-MEB to analyze the spatial transcriptome of human dorsolateral prefrontal cortex tissues and mouse hypothalamic preoptic region. Our analysis results showed that SC-MEB can achieve a similar or better clustering performance to BayesSpace, which uses the true number of clusters and a fixed smoothness parameter.

#### Sufficient Variable Screening for Ultrahigh Dimensional Right Censored Data via Independence Measures

♦*Baoying Yang<sup>1</sup>, Qingcong Yuan and Xiangrong Yin*

<sup>1</sup>Department of Statistics, College of Mathematics, Southwest Jiaotong University

yangbaoying@sjtu.edu.cn

We propose two sufficient variable screening procedures, i.e., one-stage and two-stage approaches using independence measures for ultrahigh dimensional right censored data. They are particularly useful when some active predictors are marginally independent of the response, but many existing methods fail to detect such predictors. Our procedures are model-free and thus robust against model mis-specification. We show the advantages of the proposed procedures over some existing variable screening methods through simulations and a real data analysis

#### **Mendelian Randomization Accounting for Complex Correlated Horizontal Pleiotropy While Elucidating Shared Genetic Etiology**

♦*Qing Cheng<sup>1</sup>, Xiao Zhang, Lin Chen<sup>2</sup> and Jin Liu<sup>3</sup>*

<sup>1</sup>Center of Statistical Research, School of Statistics, Southwestern University of Finance and Economics, Chengdu, Sichuan, China.

<sup>2</sup>Department of Public Health Sciences, The University of Chicago, Chicago, IL, USA.

<sup>3</sup>Centre for Quantitative Medicine, Health Services & Systems Research, Duke-NUS Medical School, Singapore, Singapore  
chengqing@swufe.edu.cn

Mendelian randomization (MR) harnesses genetic variants as instrumental variables (IVs) to study the causal effect of exposure on outcome using summary statistics from genome-wide association studies. ClassicMR assumptions are violated when IVs are associated with unmeasured confounders, i.e., when correlated horizontal pleiotropy (CHP) arises. Such confounders could be a shared gene or inter-connected pathways underlying exposure and outcome. We propose MR-CUE (MR with Correlated horizontal pleiotropy Unraveling shared Etiology and confounding), for estimating causal effect while identifying IVs with CHP and accounting for estimation uncertainty. For those IVs, we map their cis-associated genes and enriched pathways to inform shared genetic etiology underlying exposure and outcome. We apply MR-CUE to study the effects of interleukin 6 on multiple traits/diseases and identify several S100 genes involved in shared genetic etiology. We assess the effects of multiple exposures on type 2 diabetes across European and East Asian populations.

#### **A General Framework for Identifying Hierarchical Interactions and Its Application to Genomics Data**

♦*Xiao Zhang<sup>1</sup>, Xingjie Shi<sup>2</sup>, Yiming Liu<sup>3</sup>, Xu Liu<sup>3</sup> and Shuangge Ma<sup>4</sup>*

<sup>1</sup>School of Data Science, The Chinese University of Hong Kong-Shenzhen

<sup>2</sup>KLATASDS-MOE, Academy of Statistics and Interdisciplinary Sciences, East China Normal University

<sup>3</sup>School of Statistics and Management, Shanghai University of Finance and Economics

<sup>4</sup>Department of Biostatistics, Yale University  
zxiao.1994@gmail.com

The analysis of hierarchical interactions has long been a challenging problem due to the large number of candidate main effects and interaction effects, and the need for accommodating the “main effects, interactions” hierarchy. The two-stage analysis methods enjoy simplicity and low computational cost, but contradict the fact that the outcome of interest is attributable to the joint effects of multiple main factors and their interactions. The existing joint analysis methods can accurately describe the underlying data generating process, but suffer from prohibitively high computational cost. And it is not straightforward to extend their optimization algorithms to general

loss functions. To address this need, we develop a new computational method that is much faster than the existing joint analysis methods and rivals the runtimes of two-stage analysis. The proposed method, HierFabs, adopts the framework of the forward and backward stagewise algorithm and enjoys computational efficiency and broad applicability. To accommodate hierarchy without imposing additional constraints, it has newly developed forward and backward steps. It naturally accommodates the strong and weak hierarchy, and makes optimization much simpler and faster than in the existing studies. Simulations show that it outperforms the existing methods. The analysis of TCGA data demonstrates its competitive performance.

#### **Session 23CHI84: Advances in Causal Machine Learning: Challenges and Breakthrough**

##### **Matrix-Valued Network Autoregression Model with Latent Group Structure**

*Yimeng Ren<sup>1</sup>, ♦Xuening Zhu<sup>1</sup>, Ganggang Xu<sup>2</sup> and Yanyuan Ma<sup>3</sup>*

<sup>1</sup>Fudan University

<sup>2</sup>University of Miami

<sup>3</sup>The Pennsylvania State University  
xueningzhu@fudan.edu.cn

Matrix-valued time series data are frequently observed in a broad range of areas and have attracted great attention recently. In this work, we model network effects for high dimensional matrix-valued time series data in a matrix autoregression framework. To characterize the potential heterogeneity of the subjects and handle the high dimensionality simultaneously, we assume that each subject has a latent group label, which enables us to cluster the subject into the corresponding row and column groups. We propose a group matrix network autoregression (GMNAR) model, which assumes that the subjects in the same group share the same set of model parameters. To estimate the model, we develop an iterative algorithm. Theoretically, we show that the groupwise parameters and group memberships can be consistently estimated when the group numbers are correctly or possibly over-specified. An information criterion for group number estimation is also provided to consistently select the group numbers. Lastly, we implement the method on a Yelp dataset to illustrate the usefulness of the method.

##### **Sparse Causal Mediation Analysis with Unmeasured Mediator-Outcome Confounding**

*Kang Shuai<sup>1</sup>, Lan Liu<sup>2</sup>, Yangbo He<sup>1</sup> and ♦Wei Li<sup>3</sup>*

<sup>1</sup>Peking University

<sup>2</sup>University of Minnesota

<sup>3</sup>Renmin University of China  
weilistat@ruc.edu.cn

Causal mediation analysis aims to investigate how an intermediary factor called mediator regulates the causal effect of a treatment on an outcome. With the increasing availability of measurements on a large number of potential mediators in various disciplines, methods for conducting mediation analysis with many or even high-dimensional mediators have been proposed. However, they often assume there is no unmeasured confounding between mediators and the outcome. This paper allows such confounding and provides an approach to address both identification and mediator selection problems under the structural equation modeling framework. The identification strategy involves constructing a pseudo proxy variable for unmeasured confounding based on a latent factor model for multiple mediators. Using this proxy variable, we then propose a

## Abstracts

---

partially penalized procedure to select important mediators which have nonzero effects on the outcome. The resultant estimates are consistent and the estimates of nonzero parameters are asymptotically normal. Simulation studies show advantageous performance of the proposed procedure over other existing methods. We finally apply our approach to genomic data and identify gene expressions that may actively mediate the effect of a genetic variant on mouse obesity.

### Optimal Individualized Decision-Making with Proxies

♦*Tao Shen<sup>1</sup> and Yifan Cui<sup>2</sup>*

<sup>1</sup>Department of Statistics and Data Science, National University of Singapore

<sup>2</sup>Center for Data Science, Zhejiang University  
taoshen@u.nus.edu

A common concern when a policymaker draws causal inferences from and makes decisions based on observational data is that the measured covariates are insufficiently rich to account for all sources of confounding, i.e., the standard no confoundedness assumption fails to hold. The recently proposed proximal causal inference framework shows that proxy variables can be leveraged to identify causal effects and therefore facilitate decision-making. Building upon this line of work, we propose a novel optimal individualized treatment regime based on so-called outcome-inducing and treatment-inducing confounding bridges. We then show that the value function of this new optimal treatment regime is superior to that of existing ones in the literature. Theoretical guarantees, including identification, superiority, and excess value bound of the estimated regime, are established. Furthermore, we demonstrate the proposed optimal regime via numerical experiments and a real data application.

### Blessing from Human-Ai Interaction: Super Reinforcement Learning in Confounded Environments

♦*Jiayi Wang<sup>1</sup>, Zhengling Qi<sup>2</sup> and Chengchun Shi<sup>3</sup>*

<sup>1</sup>University of Texas at Dallas

<sup>2</sup>George Washington University

<sup>3</sup>London School of Economics and Political Science  
jiayi.wang2@utdallas.edu

We introduce super reinforcement learning in the batch setting, which takes the observed action as input for achieving a stronger oracle in policy learning under endogeneity. In the presence of unmeasured confounders, the recommendations from agents recorded in the observed data allow us to recover certain unobserved information. Including this information in the policy search, the proposed super reinforcement learning will yield a super-policy that is guaranteed to outperform both the standard optimal policy and the behavior one (e.g., human/AI agents' recommendations). Furthermore, to address the issue of unmeasured confounding in finding super-policies, a number of nonparametric identification are established. Based on these identification results, we develop several super-policy learning algorithms and derive their corresponding finite-sample regret guarantees. Finally, we illustrate the effectiveness of our proposal through extensive simulations and two real applications related to improving health policy.

### Session 23CHI85: Large-Scale Dependent Data Modelling and Analysis

#### Sequential One-Step Estimator by Subsampling for Customer Churn Analysis with Massive Datasets

♦*Feifei Wang<sup>1</sup>, Danyang Huang<sup>1</sup>, Tianchen Gao<sup>2</sup>, Shuyuan Wu<sup>3</sup>*

and Hansheng Wang<sup>3</sup>

<sup>1</sup>Renmin University of China

<sup>2</sup>Xiamen University

<sup>3</sup>Peking University  
feifei.wang@ruc.edu.cn

Customer churn is one of the most important concerns for large companies. Currently, massive data are often encountered in customer churn analysis, which bring new challenges for model computation. To cope with these concerns, subsampling methods are often used to accomplish data analysis tasks of large scale. To cover more informative samples in one sampling round, classic subsampling methods need to compute non-uniform sampling probabilities for all data points. However, this method creates a huge computational burden for datasets of large scale and therefore, is not applicable in practice. In this study, we propose a sequential one-step (SOS) estimation method based on repeated subsampling datasets. In the SOS method, data points need to be sampled only with uniform probabilities, and the sampling step is conducted repeatedly. In each sampling step, a new estimate is computed via one-step updating based on the newly sampled data points. This leads to a sequence of estimates, of which the final SOS estimate is their average. We theoretically show that both the bias and the standard error of the SOS estimator can decrease with increasing subsampling sizes or subsampling times. The finite sample SOS performances are assessed through simulations. Finally, we apply this

#### Transfer Learning for Spatial Autoregressive Models

♦*Hao Zeng, Wei Zhong and Xingbai Xu*

Xiamen University

wzhang@xmu.edu.cn

Political scientists use spatial models to recognize the relationship between geographic variables and the outcome of presidential elections. Traditional analysis of swing states has led to unsatisfactory predictions due to insufficiently labeled data. However, transfer learning can potentially overcome this problem by exploiting a large amount of available data from closely related states. In this paper, we present a novel transfer procedure and a two-step estimator to merge target and source models for spatial autoregressive models. This methodology has not been thoroughly investigated theoretically in the context of spatial econometrics. Furthermore, the article establishes error bounds for the proposed two-step estimator. The numerical performance of the proposed methods is investigated in a number of different settings. We apply the proposed algorithms to the exploration of the spatial relationship of the election prediction problem in the swing states using the 2016 US presidential election polling data and other demographic data with geographic information.

#### Adaptive False Discovery Rate Control with Privacy Guarantee

*Xintao Xia<sup>1</sup> and Zhanrui Cai<sup>2</sup>*

<sup>1</sup>Iowa State University

<sup>2</sup>University of Hong Kong

zhanruic@hku.hk

Differentially private multiple testing procedures can protect the information of individual hypothesis tests while guaranteeing a small fraction of false discoveries. In this paper, we propose a differentially private adaptive FDR control method that can control the classic FDR metric exactly at a user-specified level  $\alpha$  with privacy guarantee, which is a non-trivial improvement compared to the DP-BH method. Our analysis is based on two key insights: 1) a novel  $p$ -value transformation that preserves both privacy and the mirror conservative property, and 2) a mirror peeling algorithm that allows

the construction of the filtration and application of the optimal stopping technique. Numerical studies demonstrate that the proposed DP-AdaPT performs better compared to the existing differentially private FDR control methods. Compared to the original AdaPT, it only incurs a small accuracy loss but also significantly reduces the computation cost.

#### Distributed Logistic Regression for Massive Data with Rare Events

♦*Xuetong Li<sup>1</sup>, Xuening Zhu<sup>2</sup> and Hansheng Wang<sup>1</sup>*

<sup>1</sup>Peking University

<sup>2</sup>Fudan University

2001110929@stu.pku.edu.cn

Large-scale rare events data are commonly encountered in practice. To tackle the massive rare events data, we propose a novel distributed estimation method for logistic regression in a distributed system. For a distributed framework, we face the following two challenges. The first challenge is how to distribute the data. In this regard, two different distribution strategies (i.e., the RANDOM strategy and the COPY strategy) are investigated. The second challenge is how to select an appropriate type of log-likelihood function so that the best asymptotic efficiency can be achieved. Then, the under-sampled (US) and inverse probability weighted (IPW) types of log-likelihood functions are considered. Our results suggest that the COPY strategy together with the IPW log-likelihood function is the best solution for distributed logistic regression with rare events.

#### Session 23CHI86: Modern Statistics on Complex Data

##### Nsr-Ucd Model for Covid-19 Transmission with Unaware Infections.

♦*Chang Liu<sup>1</sup> and Yiyuan She<sup>2</sup>*

<sup>1</sup>Jilin University

<sup>2</sup>Florida State University  
cl18@mails.jlu.edu.cn

The mutation of COVID-19 has led to ongoing outbreaks of epidemics, which have had a major impact on human life and the economy. While existing epidemic models can be used to predict the number of COVID-19 cases, they may oversimplify the complex disease process and struggle to capture the unique features of virus transmission. This paper proposes a novel statistical model for COVID-19 transmission. The model considers individuals who are infected but unaware of their infection, as well as those who have healed without medical assistance. The model is designed with flexible assumptions regarding nonlinear infection rate and time-varying transmission rate to capture the dynamics features of COVID-19. We validated the accuracy of the model on publicly available data in Canada and found a significant number of unaware infections and deaths. Our analysis indicates that the Omicron variant has a higher recovery rate and lower death rate than the original strain. In addition, we compare the predictive performance with other commonly used methods like LSTM, ARIMA, and ES, using data from various regions.

##### Bayesian Hierarchical Model for Patient-Specific Abnormal Region Detection

*Rongjie Liu*

Florida State University  
rliu3@fsu.edu

we propose an algorithm ‘Patient-specific Abnormal Region Detection’(PARD) to identify the heterogeneous diseased regions by solving a Bayesian latent-space variable selection problem. Using

Bayesian hierarchical modeling, we account for the heterogeneity among the subjects as a large-scale variability and incorporate the inherent spatial dependence within subjects using spike-and-slab priors into the latent space. A Gibbs sampling framework is derived for estimating the model parameters and hyper-parameters in an efficient way. Simulation study shows the superiority of the proposed algorithm over popular unsupervised learning methods. The algorithm is further applied to the resting-state MRI brain scans of subjects collected from Alzheimer’s Disease Neuroimaging Initiative (ADNI) and the detected regions are validated by cross-matching with the brain’s default mode network (DMN). Additional subgroup analyses of the detected regions reveal important findings that can be helpful for developing targeted medicine and personalized treatment for Alzheimer’s patients.

##### Pivot Statistics for Normal Populations

*Liqiang Ni*

University of Central Florida  
happystat@ymail.com

Pivot statistics are widely used as basis for statistical inference, e.g. confidence interval, hypothesis testing, and predictive inference. In this expository talk, we introduce the concepts of sufficiency, pivot, invariance, and Bayesian inference. Then, we discuss the linkage between them and show some surprising results: some classic statistic long thought to be pivot is not pivot; reducing the requirement of affine invariance to low-triangular invariance can produce a pivot. We conclude with some work-in-progress in search for truepivot.

##### Consistent Dynamic Bayesian Network Learning for a Fmri Study of Emotion Processing

♦*Lizhe Sun<sup>1</sup>, Aiying Zhang<sup>2</sup> and Faming Liang<sup>3</sup>*

<sup>1</sup>Peking University

<sup>2</sup>Columbia University and New York State Psychiatric Institute

<sup>3</sup>Purdue University  
lsun@bicmr.pku.edu.cn

Learning brain connectivity networks has been an important task in the analysis of functional magnetic resonance imaging (fMRI) data, which helps people understand how brain regions cooperate with each other in a large functional network structure to handle specific cognitive processes and produce affective behavior. In this paper, we proposed a multiple conditional independence test-based method for learning dynamic Bayesian networks with high-dimensional and large-scale data. The proposed method can be regarded as a two-step method. The first step is a hybrid Bayesian integration analysis method for joint estimation of multiple graphical models from very long time series data, and the second step is the Markov neighborhood regression method for high-dimensional inference and variable selection. The proposed method is theoretically justified for its consistency and our numerical results demonstrate its robustness, providing a significant improvement over existing methods. The proposed method was implemented to investigate effective connectivity between various regions of interest (ROIs) in the brain during an emotion-processing task in a fMRI study. Our study reveals that the Subcortical-cerebellum plays a crucial role in emotion processing, and inter-modular connectivity among Subcortical-cerebellum, Motor, and Visual Association modules tend to be more active during the emotion-processing task.

##### A Sparse Ising Model with Latent Variables

*Lizhu Tao*

lizhutao@scu.edu.cn

TBC

## Abstracts

### Session 23CHI101: Advances in Analysis of Genomic and High Dimensional Data Analysis

#### Likelihood Ratio Test for Poisson Directed Acyclic Graph

Shuyan Chen<sup>1</sup>, Xin Liu<sup>2</sup>, Xiaotong Shen<sup>3</sup> and ♦Shaoli Wang<sup>2</sup>

<sup>1</sup>University of Science and Technology of China

<sup>2</sup>Shanghai University of Finance and Economics

<sup>3</sup>University of Minnesota

swang@shufe.edu.cn

Directed acyclic graphs (DAGs) are used to describe causal effects among random variables, and have wide applications in bioinformatics, neuroscience, and among others. The inference in DAGs has gained attention recently. In this talk we focus on the inference of directed linkages and directed pathways in Poisson directed graphical models. We derive the asymptotic distributions for likelihood ratio tests under both null and alternative hypotheses with non-convex acyclicity constraints in high-dimensional situations. Power analysis is given based on the asymptotic distributions. Simulations and an application are conducted for illustration.

#### Development of Pulmonary Nodule Malignancy Risk Models using Nist Data

Fenghai Duan

Brown University School of Public Health

Fenghai\_Duan@brown.edu

Lung cancer is the leading cause of cancer death worldwide each year. Non-invasive medical imaging technologies are becoming routine in screening high-risk populations for lung cancer. Unlike traditional radiological imaging analysis, which is manually interpreted by radiologists, rapid progress in computational methods and artificial intelligence has led to the extensive implementation of radiomics and/or deep learning in medical image analysis. Radiomics refers to the high-throughput extraction and analysis of quantitative features from advanced medical images with the assistance of computer science to provide a comprehensive quantification of tumor phenotype for cancer patients, while deep learning generally refers to the application of deep neural networks (such as CNNs) to process and analyze medical images. In this study, we developed and compared various models for predicting the malignancy risk of pulmonary nodules. These models were derived from semantic features, radiomic features, and artificial neural networks, using data from the National Lung Screening Trial, which is the largest lung cancer screening trial conducted to date.

#### A Statistical Learning Method for Simultaneous Copy Number Estimation and Subclone Clustering with Single Cell Sequencing Data

Fei Qin<sup>1</sup>, Guoshuai Cai<sup>1</sup> and ♦Feifei Xiao<sup>2</sup>

<sup>1</sup>University of South Carolina

<sup>2</sup>University of Florida

feifeixiao@ufl.edu

The availability of single cell sequencing (SCS) enables us to assess intra-tumor heterogeneity and identify cellular subclones without the confounding effect from mixed cells. Copy number aberrations (CNAs) have been commonly used to identify subclones in SCS data using various clustering methods since cells comprising a subpopulation are found to share genetic profile. However, currently available methods may generate spurious results (e.g., falsely identified CNAs) in the procedure of CNA detection, hence diminishing the accuracy of subclones identification from a large complex cell population. In this study, we developed a CNA detection method based on a fused lasso model, referred to as FLCNA, which can simultaneously identify subclones in single cell DNA sequencing

(scDNA-seq) data. Spike-in simulations were conducted to evaluate the clustering and CNA detection performance of FLCNA benchmarking to existing copy number estimation methods (SCOPE, HMMcopy) in combination with the existing and commonly used clustering methods. FLCNA was applied to a real scDNA-seq dataset of breast cancer with clusters identified using FLCNA showing remarkably different genomic variation patterns. In conclusion, FLCNA provided superior performance in subclone identification and CNA detection with scDNA-seq data.

#### Dna Methylation in the Eyes of a Biological Data Scientist: From Biomarkers to Functional Interpretation

Xiang Chen

St. Jude Children's Research Hospital

xiang.chen@stjude.org

Pediatric cancers seldom have strong environmental image and are essentially a disease of deregulated development. Pediatric cancer genome project and other cancer genomic studies consistently revealed that pediatric cancer harbors few genetic drivers. On the other side, epigenetic regulations play a critical role in both physiological and pathological development. However, limited availability of high quality tissues for ChIP-seq profiling and lack of functional interpretation of DNA methylome changes remain as major roadblocks to decipher the epigenetic drivers in pediatric tumors. We have developed a deep-learning based approach to interpret the functional consequence of DNA methylation changes. I will further demonstrate the functionality in a real biological application of rhabdomyosarcomas.

### Session 23CHI102: Frontiers in -Omics Data Analysis

#### Analysis of Spatial Transcriptomics using Optimal Transport

Zixuan Cang

North Carolina State University

zcang@ncsu.edu

The emerging single-cell and spatial genomics techniques allow us to elucidate the governing rules of multicellular systems with unprecedented resolution and depth. These datasets are often high-dimensional, complex, and heterogeneous. Mathematical tools are needed to extract biological insights from such data. In this talk, we will discuss several computational methods for exploring tissue structures, temporal signatures, and cell-cell communication processes on spatial transcriptomics data as well as supervised optimal transport motivated by the biological applications.

#### Spatially Aware Dimension Reduction for Spatial Transcriptomics

♦Lulu Shang and Xiang Zhou

University of Michigan

shanglu@umich.edu

Spatial transcriptomics are a collection of genomic technologies that have enabled transcriptomic profiling on tissues with spatial localization information. Analyzing spatial transcriptomic data is computationally challenging, as the data collected from various spatial transcriptomic technologies are often noisy and display substantial spatial correlation across tissue locations. Here, we develop a spatially-aware dimension reduction method, SpatialPCA, that can extract a low dimensional representation of the spatial transcriptomics data with biological signal and preserved spatial correlation structure, thus unlocking many existing computational tools previously developed in single-cell RNAseq studies for tailored and novel analysis of spatial transcriptomics. We illustrate the benefits

of SpatialPCA for spatial domain detection and explore its utility for trajectory inference on the tissue and for high-resolution spatial map construction. In real data applications, SpatialPCA identifies key molecular and immunological signatures in a newly detected tumor surrounding microenvironment, including a tertiary lymphoid structure that shapes the gradual transcriptomic transition during tumorigenesis and metastasis. In addition, SpatialPCA detects the past neuronal developmental history that underlies the current transcriptomic landscape across tissue locations in the cortex.

#### Spatial Dependency-Aware Deep Generative Models

♦Tian Tian<sup>1</sup>, Jie Zhang<sup>2</sup>, Xiang Lin<sup>2</sup>, Zhi Wei<sup>2</sup> and Hakon Hakonarson<sup>1</sup>

<sup>1</sup>Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA.

<sup>2</sup>Department of Computer Science, New Jersey Institute of Technology, Newark, New Jersey, USA.  
tt72@njit.edu

With the advances of single-cell and spatial sequencing techniques, analysis has been touted for diverse biomedical questions. However, the analysis of single-cell data remains computationally and analytically challenging due to the discrete, over-dispersed and degree of noise in the data. For the spatial genomics data, the situation is further compromised by complex spatial dependencies. To address these limitations, we propose model-based deep learning approaches for various analysis of single-cell and spatial genomics data. First, we will discuss a model-based deep hyperbolic manifold learning approach to visualize complex hierarchical structures in single-cell genomics data. Second, we will propose dependency-aware deep generative model for multitasking analysis of spatial genomics data, including dimensionality reduction, visualization, clustering, batch integration, denoising, differential expression, spatial imputation, resolution enhancement, and identifying spatial genes.

#### Genome-Wide Prediction of Structural Variations and Enhancer-Hijacking Events from Chromatin Interaction Data in Cancer Genomes

♦Xiaotao Wang<sup>1</sup>, Jie Xu<sup>2</sup>, Baozhen Zhang<sup>3</sup>, Yu Luan<sup>4</sup>, Fan Song<sup>5</sup>, Hongbo Yang<sup>1</sup>, Juan Wang<sup>4</sup>, Tingting Liu<sup>6</sup> and Feng Yue<sup>4</sup>

<sup>1</sup>Fudan University

<sup>2</sup>University of California San Diego

<sup>3</sup>Peking University Cancer Hospital and Institute

<sup>4</sup>Northwestern University

<sup>5</sup>Illumina

<sup>6</sup>Southeast University

wangxiaotao@fudan.edu.cn

Recent research has shown that structural variations (SVs) can disrupt three-dimensional genome organization and induce enhancer hijacking; however, there are currently no computational tools available to identify such events using chromatin interaction data. To address this gap, we have developed two novel computational methods. The first is EagleC, which combines deep-learning and ensemble-learning techniques to accurately predict a full range of SVs from chromatin interaction data. The second is NeoLoopFinder, which is a framework that identifies novel chromatin interactions resulting from SVs. Our methods can handle complex SVs, reconstruct local Hi-C maps around SV breakpoints, normalize copy number variation and allele effects, and predict chromatin loops induced by SVs. Using these methods, we have successfully identified instances of enhancer hijacking in primary tumors from various cancer types, such as bladder cancer, medul-

loblastoma, acute myeloid leukaemia, and liposarcoma. To validate our findings, we used CRISPR-Cas9 and CRISPRi to delete or disrupt the hijacked enhancers, which led to a significant reduction in the expression of the target oncogenes. In summary, our methods provide a means of identifying oncogenic regulatory elements, which could potentially be used to identify novel therapeutic targets.

#### Session 23CHI156: Mitigating Incomplete Data Biases-a Modern Take

##### Robust and Efficient Case-Control Studies with Contaminated Case Pools: a Unified m-Estimation Framework

♦Guorong Dai<sup>1</sup> and Jinbo Chen<sup>2</sup>

<sup>1</sup>Fudan University

<sup>2</sup>University of Pennsylvania  
guorongdai@fudan.edu.cn

We consider a general M-estimation problem based on contaminated case-control data, including the primary and secondary analyses of case-control studies as special examples. The case pool contains ineligible patients who should be excluded from the study if known, but the true status of an individual in the case pool is unclear except in a small subset. Through imputing the possibly unobserved status variable with a function of all available relevant predictors, followed by an appropriate debiasing procedure, we exploit the whole sample to develop a family of robust and efficient estimators, eliminating bias from the case contamination. The imputation function can be constructed using any reasonable regression or machine learning approaches. Our estimators are always root-n-consistent and asymptotically normal regardless of the imputation function's limit. Further, we explore relaxation of requirements on the imputation function. We show even without any assumption on its convergence properties, a cross-fitting version of our estimators is still root-n-consistent while asymptotic normality can be achieved by a sample-splitting variant. The finite-sample superiority of our method is demonstrated by comprehensive simulation studies. We also apply our method to analyze sepsis-related death based on a real data set from electronic health records.

#### Session 23CHI163: New Generation of Statisticians in Drug Development

#### Session 23CHI167: Causal Inference: From Practice to Theory

##### Inference of Treatment Effects and Complier Treatment Effects under Computer Assisted Balance-Improving Designs

♦Junni Zhang<sup>1</sup>, Per Johansson<sup>2</sup> and Zhen Zhong<sup>3</sup>

<sup>1</sup>National School of Development, Peking University

<sup>2</sup>Department of Statistics, Uppsala University

<sup>3</sup>Department of Mathematical Sciences, Tsinghua University  
zjn@nsd.pku.edu.cn

To improve covariate balance over a complete randomization, a number of methods have been proposed recently to utilize modern computational capabilities to find allocations with balance in observed covariates. In experiments based on these designs, asymptotic inference on treatment effects when there is perfect compliance or on complier treatment effects when there is noncompliance is much more complicated than that under complete randomization. When there is perfect compliance, we focus on estimating the Sample Average Treatment Effect (SATE), and propose two approaches

## Abstracts

---

for inference: regression adjustment together with small sample adjusted estimators of standard errors and model-based Bayesian approach. When there is noncompliance, we focus on estimating the Sample Complier Average Treatment Effect (SCATE), and also propose two approach for inference: a nonparametric estimator and model-based Bayesian approach. We examine the small sample performance of the approaches using Monte Carlos simulations with the Mahalanobis based rerandomization in the experimental design.

### **Interpretable Sensitivity Analysis for the Baron – kenny Approach to Mediation with Unmeasured Confounding**

*Mingrui Zhang and ♦Peng Ding*

University of California Berkeley

[pengdingpku@berkeley.edu](mailto:pengdingpku@berkeley.edu)

Mediation analysis assesses the extent to which the treatment affects the outcome indirectly through a mediator and the extent to which it operates directly through other pathways. As the most popular method in empirical mediation analysis, the Baron–Kenny approach estimates the indirect and direct effects of the treatment on the outcome based on linear structural equation models. However, when the treatment and the mediator are not randomized, the estimates may be biased due to unmeasured confounding among the treatment, mediator, and outcome. Building on Cinelli and Hazlett (2020a), we propose a sharp and interpretable sensitivity analysis method for the Baron – Kenny approach to mediation in the presence of unmeasured con- founding. We first modify their omitted-variable bias formula to facilitate the discussion with heteroskedasticity and model misspecification. We then apply the result to develop a sensitivity analysis method for the Baron–Kenny approach. To ensure interpretability, we express the sensitivity parameters in terms of the partial R<sup>2</sup>'s that correspond to the natural factorization of the joint distribution of the direct acyclic graph for mediation analysis. They measure the proportions of variability explained by unmeasured confounding given the observed variables. Moreover, we extend the method to deal with multiple mediators, based on a

### **Optimal Treatment Regime under the Individual No-Harm Criterion**

*Peng Wu*

Beijing Technology and Business University

[pengwu@btbu.edu.cn](mailto:pengwu@btbu.edu.cn)

Trustworthy optimal treatment regime has significant importance in making trustworthy and harmless treatment decisions for individuals. Previous treatment regime estimation approaches aim at the well-being of subgroups by maximizing the utility function (e.g., conditional average causal effects), however, individual-level counterfactual no-harm criterion has rarely been discussed. In this paper, we first formalize the counterfactual no-harm criterion for treatment regime from a principal stratification perspective. Next, we propose a novel upper bound for the fraction negatively affected by the treatment regime and show the consistency and asymptotic normality of the estimator. Based on the estimators for the treatment regime utility and harm upper bounds, we further propose a treatment regime estimation approach that satisfies the counterfactual no-harm criterion, and prove its consistency to the optimal no-harm treatment regime reward for parametric and nonparametric class of deterministic decision rules, respectively. Extensive experiments are conducted to show the effectiveness of the proposed treatment regime estimation approach for satisfying the counterfactual no-harm criterion.

### **Session 23CHI174: Special Memorial Session to Celebrate Life of Professor Tze Leung Lai**

#### **Special Memorial Session to Celebrate Life of Professor Tze Leung Lai**

*Ka Wai Tsang*

The Chinese University of Hong Kong at Shenzhen

[kwtsang@cuhk.edu.cn](mailto:kwtsang@cuhk.edu.cn)

Since I became Professor Lai's PhD student at Stanford in 2011, Professor Lai had helped me a lot in various aspects of my life. In this session, I will give a brief introduction of some Lai's significant work during the time I was learning from him at Stanford. Besides, I will also share my experience of working with Professor Lai.

#### **Best Arm Identification in Batched Multi-Armed Bandit Problems**

*Shengyu Cao, Simai He, Jin Xu and ♦Hongsong Yuan*

Shanghai University of Finance and Economics

[yuan.hongsong@shufe.edu.cn](mailto:yuan.hongsong@shufe.edu.cn)

Recently multi-armed bandit problem arises in many real-life scenarios where arms must be sampled in batches due to limited time the agent can wait for the feedback. Such applications include biological experimentation and online marketing. The problem is further complicated when the number of arms is large. We consider pure exploration in a batched multi-armed bandit problem with where the number of arms tends to infinity, and the number of batches is relatively small. We introduce a linear programming framework that can serve as a proxy for the optimal algorithm to identify the best arm. The linear program naturally leads to a two-step algorithm that can achieve good theoretical properties. We demonstrate by simulation that the algorithm also has good performance in terms of both regret and total arm pulls.

#### **Causal IV Analysis for Time-to-Event Data**

*Gang Li*

UCLA

[vli@ucla.edu](mailto:vli@ucla.edu)

Professor Lai was an exceptional scholar, a globally recognized leader, and an esteemed educator. Professor Lai's groundbreaking work encompasses diverse areas such as sequential statistical analysis, the multi-armed bandit problem, stochastic approximation and recursive estimation, adaptive control of linear stochastic systems and Markov decision processes, saddlepoint approximations, boundary-crossing probabilities in Markov random walks and random fields, survival analysis, and inference for hidden Markov models. In this talk, I will present some recent research aimed at tackling some computational and statistical challenges particularly in causal inference for an event time outcome involved in analyzing massive electronic health records (EHRs) and biobank data. These endeavors have been inspired by the earlier work of Professor Lai and his students in the realm of survival analysis.

#### **深切怀念黎先生**

*Dong Han*

Shanghai Jiaotong University

[donghan@sjtu.edu.cn](mailto:donghan@sjtu.edu.cn)

一是缅怀黎先生对祖国的拳拳之心；二是回忆黎先生对上海交大统计学科的支持以及对我本人的指导和帮助。

## Session 23CHI58: From Early Development to PAC: Application of Bayesian Analysis in Drug Development

### A Bayesian Sample Size Planning Tool for Phase i Dose-Finding Trials

Xiaolei Lin

Fudan University

xiaoleilin@fudan.edu.cn

We propose BaySize, a sample size calculator for phase I clinical trials using Bayesian models. BaySize applies the concept of effect size in dose finding, assuming the MTD is dened based on an equivalence interval. Leveraging a decision framework that involves composite hypotheses, BaySize utilizes two prior distributions, the fitting prior (for model fitting) and sampling prior (for data generation), to conduct sample size calculation under desirable statistical power. Look-up tables are generated to facilitate practical applications. To our knowledge, BaySize is the first sample size tool that can be applied to a broad range of phase I trial designs.

### Application of Bayesian Analysis in Early Phase Drug Development

Chengyuan Song

boehringer-ingelheim

songch1990@gmail.com

Early phase trials are an integral component of the development plan and constitute the generation of data that provide reassurance regarding key elements such as proof of principle, proof of concept and dose selection. There is an ever-increasing number of early phase clinical designs and study analyses using Bayesian frameworks. Bayesian statistics provide a formal mathematical method for combining prior information with current information at the design stage. There are many advantages of Bayesian statistics in early phase designs: reduce sample size, make the study more flexible, borrow information between cohorts and expedite drug development. The presentation will focus on the application of Bayesian analysis in early phase drug development: use of Bayesian approaches to guide the dose escalation, the dose optimization and dose selection.

### Basic: a Bayesian Adaptive Synthetic Control Design for Phase Ii Clinical Trials

♦Liyun Jiang, Peter F Thall, Fangrong Yan, Scott Kopetz and Ying Yuan  
3253727485@qq.com

Randomized controlled trials (RCTs) are considered the gold standard for evaluating experimental treatments, but often require large sample sizes. Single-arm trials (SA) require smaller sample sizes, but are subject to bias when using historical control data (HCD) for comparative inferences. This article presents a Bayesian adaptive synthetic control (BASIC) design that exploits HCD to create a hybrid of a single-arm trial and an RCT. BASIC has two stages. In stage 1, a prespecified number of patients are enrolled in a single arm given the experimental treatment. Based on the stage 1 data, applying propensity score matching and Bayesian posterior prediction methods, the usefulness of the HCD for identifying a pseudo-sample of matched synthetic control patients for making comparative inferences is evaluated. If a sufficient number of synthetic controls can be identified, the single arm trial is continued. If not, the trial is switched to an RCT. BASIC achieves power and unbiasedness similar to an RCT, but on average requires a much smaller sample size, provided that HCD patients are sufficiently comparable to trial patients so that a good number of matched controls can be identified in

the HCD. Compared to SA, BASIC yields much higher power and much smaller bias.

### Bayesian Dynamic Borrowing Method Implementation in a Phase III Post-Marketing Trial

Ling Li

ling\_3.li@boehringer-ingelheim.com

In December 2020, Nintedanib 150 mg bid was approved in China for the treatment of chronic fibrosing Interstitial Lung Diseases with a progressive phenotype (PF-ILD) with a post-marketing required randomized, placebo-controlled trial (1199-0434, NCT05065190). The aim of this trial is to explore the Nintedanib treatment effect and show directional consistency in efficacy and safety in Chinese PF-ILD patients, this study was not powered for formal hypothesis testing. To provide more reliable treatment effect estimation, Bayesian dynamic borrowing method use existing trial data from Nintedanib development program was planned with respect to the primary endpoint, annual rate of decline in Forced Vital Capacity (FVC) over 52 weeks.

## Session 23CHI60: Novel Machine Learning Methods to Advance Precision Medicine using Big Biomarker Data

### A Flexible Summary-Based Colocalization Method with Application to the Mucin Cystic Fibrosis Lung Disease Modifier Locus

♦Fan Wang<sup>1</sup>, Naim Panjwani<sup>2</sup>, Cheng Wang<sup>2</sup>, Lei Sun<sup>3</sup> and Lisa Strug<sup>3</sup>

<sup>1</sup>Columbia University

<sup>2</sup>The Hospital for Sick Children

<sup>3</sup>University of Toronto  
fw2400@cumc.columbia.edu

Integrating omics data using colocalization methods can help us understand how Genome-wide association study (GWAS) loci contribute to disease. However, contemporary colocalization analytical tools need to be powerful and valid under various complex scenarios, including overlapping samples, meta-analyses, high linkage equilibrium, and allelic heterogeneity. We previously developed the Simple Sum (SS), a powerful colocalization test in regions with allelic heterogeneity, but SS assumed eQTLs to be present to achieve type I error control. Here we propose a two-stage SS (SS2) colocalization test that avoids a priori eQTL assumptions, accounts for multiple hypothesis testing and the composite null hypothesis and enables meta-analysis. We compare SS2 to published approaches through simulation and demonstrate type I error control for all settings with the greatest power in the presence of high LD and allelic heterogeneity. Applying SS2 to the MUC20/MUC4 CF lung disease locus with eQTLs from CF HNE revealed significant colocalization with MUC4 ( $p = 1.31 \times 10^{-5}$ ) rather than with MUC20. The SS2 is a powerful method to inform the responsible gene(s) at a locus and guide future functional studies. SS2 has been implemented in the application LocusFocus (locusfocus.research.sickkids.ca).

### Blockwise Mixed Membership Models for Multivariate Longitudinal Categorical Data

♦Kai Kang<sup>1</sup> and Yugi Gu<sup>2</sup>

<sup>1</sup>Sun Yat-sen University

<sup>2</sup>Columbia University  
kangk5@mail.sysu.edu.cn

Mixed membership modeling is a popular framework for multivariate and/or longitudinal data where each individual is represented as

## Abstracts

---

a unique mixture across clusters. We develop a new class of block-wise mixed membership models that partition multivariate variables into groups and divide longitudinal observations into time-periods. The combination of each group and period forms a block such that the latent membership is same for data within a block and can be different across blocks. To further address individual heterogeneity, the structure of time-periods is assumed different across individual subgroups. Markov Chain Monte Carlo (MCMC) algorithms with efficient Gibbs sampler method are adopted for Bayesian inference. Simulation results confirm that the performance of the proposed methods is satisfactory. The proposed method is applied to a real dataset about Parkinson's disease to learn the mechanism of disease progression.

### **Knockoff-Based Statistics for the Identification of Putative Causal Genes in Genetic Studies**

♦*Shiyang Ma<sup>1</sup>, Linxi Liu<sup>2</sup>, Zihuai He<sup>3</sup> and Iuliana Ionita-Laza<sup>4</sup>*

<sup>1</sup>Shanghai Jiao Tong University School of Medicine

<sup>2</sup>University of Pittsburgh

<sup>3</sup>Stanford University

<sup>4</sup>Columbia University  
mashiyang1991@sjtu.edu.cn

Gene-based tests are important tools for elucidating the genetic basis of complex traits. Despite substantial recent efforts in this direction, the existing tests are still limited, owing to low power and detection of false-positive signals due to the confounding effects of linkage disequilibrium. In this talk, we describe a gene-based test that attempts to address these limitations by incorporating data on long-range chromatin interactions, several recent technical advances for region-based testing, and the knockoff framework for synthetic genotype generation. Through extensive simulations and applications to multiple diseases and traits, we show that the proposed test increases the power over state-of-the-art gene-based tests and provides a narrower focus on the possible causal genes involved at a locus. We will also discuss a computationally efficient gene-based testing approach for biobank-scale data, and show applications to UK Biobank data with 405,296 participants for multiple binary and quantitative traits.

### **Longitudinal Classification and Forecast in the Absence of True Disease Labels**

♦*Zexi Cai<sup>1</sup>, Karen Marder<sup>2</sup>, Donglin Zeng<sup>3</sup> and Yuanjia Wang<sup>1</sup>*

<sup>1</sup>Department of Biostatistics, Columbia University

<sup>2</sup>Department of Neurology, Columbia University

<sup>3</sup>Department of Biostatistics, University of North Carolina at Chapel Hill  
zc2626@columbia.edu

Classification and forecasts of patients' disease progression using time-series sequence data or unbalanced longitudinal data are difficult tasks since the health assessments are typically temporally dependent, where each observation is correlated with previous and future observations. This dependence makes it challenging to identify the underlying patterns and important features of disease status. In addition, the absence of gold-standard neurological disease diagnoses due to a lack of objective biomarkers further poses challenges to accurate classifications. Although surrogate labels are sometimes available by expert evaluation, they are subjective and may lead to misclassifications if used in a naive way. With available feature variables informative of the true disease diagnostic class, we propose an integration of the hidden Markov model as a generative model and a discriminative model to achieve the goal efficiently. A multinomial logistic implements the latter model and can

serve as a standalone classifier for future patients. The Expectation-Maximization algorithm is utilized to handle the missing labels, and algorithms are developed for forecasting future disease progression. Simulation studies show promising finite sample performance. We apply the proposal to data from the National Alzheimer's Coordinating Center (NACC) to distinguish Alzheimer's disease (AD) from AD and related dementias (ADRD).

## **Session 23CHI63: Technical Advances on Analyzing Single Cell RNA-Seq and Spatial Transcriptomic Data**

### **Intelligent Spatial Transcriptomics: Paving the Way for Deciphering Tissue Architecture**

*Shihua Zhang  
zsh@amss.ac.cn*

Technological advances in spatial transcriptomics are critical for a better understanding of the structure and function of tissues in biological research. Recently, the combination of intelligent/statistical algorithms and spatial transcriptomics are emerging to pave the way for deciphering tissue architecture. In this talk, I will introduce our efforts to advance intelligent spatial transcriptomics. We first develop a graph attention auto-encoder framework STAGATE to accurately identify spatial domains by learning low-dimensional latent embeddings via integrating spatial information and gene expression profiles. We validate STAGATE on diverse spatial transcriptomics datasets generated by different platforms with different spatial resolutions. STAGATE could substantially improve the identification accuracy of spatial domains, and denoise the data while preserving spatial expression patterns. Importantly, STAGATE could be extended to multiple consecutive sections to reduce batch effects between sections and extract three-dimensional (3D) expression domains from the reconstructed 3D tissue effectively. Based on this, we 1) develop STAMarker for identifying spatial domain-specific variable genes, 2) design STAligner for integrating spatial transcriptomics of multiple slices from diverse biological scenarios, and 3) illustrate the effectiveness of the graph attention auto-encoder for spatial clustering of spatial metabolomics.

### **Identifying Phenotype-Associated Subpopulations by Integrating Bulk and Single-Cell Sequencing Data**

*Duanchen Sun  
dcsun@sdu.edu.cn*

Single-cell RNA sequencing distinguishes cell types, states, and lineages within the context of heterogeneous tissues. However current single-cell data cannot directly link cell clusters with specific phenotypes. Here we present Scissor, a method that identifies cell subpopulations from single-cell data that are associated with a given phenotype. Scissor integrates phenotype-associated bulk expression data and single-cell data by first quantifying the similarity between every single cell and each bulk sample. It then optimizes a regression model on the correlation matrix with the sample phenotype to identify relevant subpopulations. Applied to a lung cancer single-cell RNA-seq dataset, Scissor identified subsets of cells associated with worse survival and with TP53 mutations. In melanoma, Scissor discerned a T cell subpopulation with low PDCD1/CTLA4 and high TCF7 expression associated with an immunotherapy response. Beyond cancer, Scissor was effective in interpreting Facioscapulohumeral muscular dystrophy and Alzheimer's disease datasets. Scissor identifies biologically and clinically relevant cell subpopulations from single-cell assays by leveraging phenotype and bulk-omics datasets.

## **Integrative Models of Single Cell Genomics Data for Dissecting Cellular Heterogeneity and Transcriptional Regulation**

*Lihua Zhang*

Wuhan University

[zhanglh@whu.edu.cn](mailto:zhanglh@whu.edu.cn)

The number and the variety of single cell genomics datasets have grown tremendously in recent years. Particularly, an increasing number of single cell datasets have been collected under different biological conditions, e.g., control vs. perturbation. Distinguishing biological from technical variation is crucial when integrating and comparing single cell genomics datasets across different experiments. Moreover, the emergence of single cell multi-omics data provides unprecedented opportunities to scrutinize the transcriptional regulatory mechanisms controlling cell identity. In this talk, I will introduce our efforts in single cell genomics data integration for dissecting cellular heterogeneity and transcriptional regulation code.

## **Spider: a Flexible and Unified Framework for Simulating Spatial Transcriptomics Data**

*Xiaoqi Zheng*

Shanghai Jiao Tong University School of Medicine

[xqzheng@shsmu.edu.cn](mailto:xqzheng@shsmu.edu.cn)

Spatial transcriptomics technology provides a valuable view for studying cellular heterogeneity due to its ability to simultaneously acquire gene expression profile and cell location information. However, benchmarking the rapidly accumulating spatial transcriptomics analysis algorithms is challenging, as there is a limited number of manually annotated “gold standard” data sets available, which lacks of diversity and accuracy. To address this issue, we proposed Spider, a flexible and unified simulator for spatial transcriptomics data guided by cell type proportion and transition matrix of nearby cell types. Taking advantage of a heuristic batched simulated annealing algorithm (BSA) in assigning the simulated cell type labels, Spider can generate spatial transcriptomics data for one million cells in just five minutes. Spider can also generate a variety types of spatial transcriptome data ranging from hot/cold tumor samples by specifying different input prior proportions and transition matrices, and enable generating layered tissue samples through an interactive mode. In addition, Spider is also a unified framework for spatial transcriptome simulation for its ability to include various simulators as special cases. Using simulated data from Spider and other tools, we have systematically summarized and compared current spatial transcriptomics data simulation algorithms. Further details and code are available at <https://github.com/YANG-ERA/Artist>.

## **Session 23CHI64: Model Estimation and Hypothesis Testing of High Dimensional Data**

### **Use of Random Integration to Test Equality of High Dimensional Covariance Matrices**

♦*Jiang Yunlu*<sup>1</sup>, *Wen Canhong*<sup>2</sup>, *Jiang Yukang*<sup>3</sup>, *Wang Xueqin*<sup>2</sup> and *Zhang Heping*<sup>4</sup>

<sup>1</sup>Jinan University

<sup>2</sup>University of Science and Technology of China

<sup>3</sup>Sun Yat-Sen University

<sup>4</sup>Yale University

[tjiangyl@jnu.edu.cn](mailto:tjiangyl@jnu.edu.cn)

Testing the equality of two covariance matrices is a fundamental problem in statistics, and especially challenging when the data are high dimensional. By means of a novel use of random integration,

we test the equality of high-dimensional covariance matrices without assuming parametric distributions for the two underlying populations, even if the dimension is much larger than the sample size. The asymptotic properties of our test for an arbitrary number of covariates and sample size are studied in depth under a general multivariate model. The finite-sample performance of our test is evaluated using numerical studies. The empirical results demonstrate that the proposed test is highly competitive with existing tests in a wide range of settings, and particularly powerful when there exist a few large or many small diagonal disturbances between the two covariance matrices.

### **Robust Optimal Subsampling Based on Weighted Asymmetric Least Squares**

*Min Ren*<sup>1</sup>, *Shengli Zhao*<sup>1</sup>, ♦*Mingqiu Wang*<sup>1</sup> and *Xinbei Zhu*<sup>2</sup>

<sup>1</sup>Qufu Normal University

<sup>2</sup>Virginia Tech University

[mqwang@vip.126.com](mailto:mqwang@vip.126.com)

With the development of contemporary science, a large amount of generated data includes the contamination in the response and/or covariates. Furthermore, subsampling is an effective method to overcome the limitation of computational resources. However, when data include heterogeneity, outliers and contamination, incorrect subsampling probabilities may select inferior subdata, and statistic inference on this subdata may have a far inferior performance. This paper proposes a Double-Robustness Framework (DRF), which is robust to both contamination and outliers using the asymmetric least squares and L2 estimation for massive data. The Poisson subsampling is implemented based on the DRF, and a more robust probability will be derived to select the subdata. Under some regularity conditions, we establish the asymptotic properties of the subsampling estimator based on the DRF. Numerical studies and actual data demonstrate the effectiveness of the proposed method.

### **A Scale-Invariant Test on General Linear Hypothesis in High-Dimensional Heteroscedastic One-Way Manova**

♦*Mingxiang Cao*<sup>1</sup> and *Ziyang Cheng*<sup>2</sup>

<sup>1</sup>Anhui Normal University

<sup>2</sup>Changchun University of Technology  
[caomingx@163.com](mailto:caomingx@163.com)

For the general linear hypothesis testing problem in high-dimension setting, several interesting tests have been proposed in the literature. In this paper, a new scale invariant test for general linear hypothesis testing problem is proposed and studied in high-dimension setting. Different from the existing test statistics which impose strong assumptions on the underlying group covariance matrices so that their test statistics are asymptotically normal, it is shown that our test statistic under some regularity conditions and null hypothesis have the same normal or non-normal limiting distributions with a chi-square-type mixture. The distribution of the chi-square-type mixture can be well approximated by the Welch-Satterthwaite chi-square-approximation with the approximation parameter consistently estimated from the data. The performance of the proposed test is conducted by simulation, which illustrates our new test outperforms competitors in the considered cases.

### **Relative Error Model Average**

♦*Xiaochao Xia*<sup>1</sup>, *Hao Ming*<sup>1</sup> and *Jialiang Li*<sup>2</sup>

<sup>1</sup>Chongqing University

<sup>2</sup>National University of Singapore  
[xxc@cqu.edu.cn](mailto:xxc@cqu.edu.cn)

We propose a relative error model averaging (REMA) approach to predict positive response values under a set of multiplicative error

## Abstracts

---

models. To estimate the parameters in each candidate multiplicative model, we utilize a relative error loss as the empirical objective function. Specifically, we consider two commonly used loss functions: the least product relative error (LPRE) and the least absolute relative error (LARE), under which two model averaging estimators, REMA-LPRE and REMA-LARE, are proposed accordingly. The optimal weight vector,  $w$ , is chosen by minimizing a jackknife version of the relative error loss. Theoretically, it is shown that under some technical conditions, our proposed model averaging estimators enjoy asymptotic optimality under the two losses, respectively, in the sense that its loss defined by a final prediction error (FPE) is asymptotically identical to that of the best yet infeasible model averaging estimator. Furthermore, we propose a model-based screening approach to deal with the high-dimensional data setting when the number of candidate models are extremely large, and then present an extension to relax the sum-to-one constraint. Extensive simulations and empirical applications are conducted to demonstrate the practical performance of our approach.

### Session 23CHI87: Advances in Machine Learning with Applications

#### Properties of Standard and Sketched Kernel Fisher Discriminant

Heng Lian  
shellinglianhang@hotmail.com

Kernel Fisher discriminant (KFD) is a popular tool as a nonlinear extension of Fisher's linear discriminant, based on the use of the kernel trick. However, its asymptotic properties are still rarely studied. We first present an operator-theoretical formulation of KFD which elucidates the population target of the estimation problem. Convergence of the KFD solution to its population target is then established. However, the complexity of finding the solution poses significant challenges when  $n$  is large and we further propose a sketched estimation approach based on a  $m \times n$  sketching matrix which possesses the same asymptotic properties (in terms of convergence rate) even when  $m$  is much smaller than  $n$ .

#### Skewed Pivotal-Point-Adaptive Modeling with Applications to Semicontinuous Outcomes

♦Yiyuan She, Xiaoqiang Wu, Lizhu Tao and Debajyoti Sinha  
yshe@stat.fsu.edu

Skewness is a common occurrence in statistical applications. In recent years, various distribution families have been proposed to model skewed data by introducing unequal scales based on the median or mode. However, we argue that the point at which unbalanced scales occur may be at any quantile and cannot be reparametrized as an ordinary shift parameter in the presence of skewness. In this paper, we construct a new skewed density family from any given continuous density, which may be asymmetric and nonunimodal, and propose a new method to simultaneously estimate the scales, pivotal point, and other location parameters. Our framework has numerous extensions and can be applied to skewed two-part models with joint variable selection. Our theoretical analysis reveals the influence of skewness without assuming asymptotic conditions. Experiments on synthetic and real-life data demonstrate the excellent performance of the proposed method.

#### when Mediation Analysis Faces Subgroup Heterogeneity

Kaizhou Lei<sup>1</sup>, Shengxian Ding<sup>1</sup>, Lexin Li<sup>2</sup>, Rongjie Liu<sup>1</sup> and ♦Chao Huang<sup>1</sup>

<sup>1</sup>Florida State University

<sup>2</sup>UC Berkley  
chaohuang.stat@gmail.com

Mediation analysis is an essential tool in the imaging genetics study for Alzheimer's disease (AD). The goal is to identify the causal mechanism or pathway that links genetic exposures and neurological outcomes through some neuroimaging mediators. Although various mediation analysis approaches have been proposed to discover the underlying causal pathway in AD, there are several challenges, such as the subgroup heterogeneities in terms of (i) brain connectomes and (ii) causal mechanisms. To address these issues, we propose a novel mediation analysis tool that can simultaneously detect individual brain connectomes and subgroup causal pathways. Specifically, a two-layer structure equation model, including a mixture of conditional Gaussian graphical models, is developed to establish the heterogeneous mediation pathways. A penalized EM algorithm is proposed to estimate both average direct effect and indirect effect. Both simulation studies and a real example analysis using the diffusion tensor imaging data from both ADNI and HCP studies are conducted to assess the finite sample performance of our method.

#### Single Index Models with Regularized Matrix Coefficients

Luo Xiao

North Carolina State University  
lxiao5@ncsu.edu

Single index models with regularized matrix coefficients Single index models extend standard linear models to account for nonlinearity between multivariate predictors and responses. We study single index models where the unknown coefficients can be formulated as a matrix and enforce regularization term(s) on the coefficient matrix to induce meaningful structure, e.g., sparsity and low-rank. We propose an iterative estimation procedure in which an alternating direction method of multipliers (ADMM) algorithm is employed to accommodate multiple regularization terms. We focus on two particular models: scalar response on matrix predictor model and multivariate response on multivariate predictor model. We apply the former model to study nonlinear association between functional connectivity networks and fluid intelligence, and the latter model to a genetic association study. The work is based on two papers, "Sparse single index models for multivariate responses" which was published in Journal of Computational and Graphical Statistics and "Single index models with functional connectivity network predictors", which was published in Biostatistics.

### Session 23CHI90: Advances in Statistical Methodologies for Clustered Data Analysis and Clustered Randomized Trials Design

#### Estimation and Inference for Complexly Correlated Data via Network Generalized Estimating Equations

♦Tom Chen<sup>1</sup>, Fan Li<sup>2</sup> and Rui Wang<sup>1</sup>

<sup>1</sup>Harvard Medical School

<sup>2</sup>Yale University  
tomchen00@gmail.com

Increasingly complex variations of cluster randomized trials (CRTs) are being developed and implemented for pragmatic trials. For example, a 3-level hierarchical design may consider interventions randomly assigned to practice, and providers within the same practice are then instructed to administer that intervention. In another context, a stepped wedge CRT (SW-CRT) randomizes the time periods

for which intact clusters of individuals crossover from control to intervention until all have been exposed. These two examples underscore the increasingly complex dependency structures which arises in clustered data settings and motivate a unifying framework to not only represent these structures, but to also tractably estimate them. The purpose of this talk is twofold: we introduce a framework to express highly complex dependency structures through the use of network theory and propose efficient estimation via a generalized estimating equations (GEE). With this flexible representation, our approach naturally facilitates robust estimation and inference for the correlation parameters which characterize these structures. Furthermore, our algorithmic implementation addresses computational issues which normally afflict GEE.

#### **Marginal Proportional Hazards Models for Clustered Interval-Censored Data with Time-Dependent Covariates**

♦*Kaitlyn Cook*<sup>1</sup>, *Wenbin Lu*<sup>2</sup> and *Rui Wang*<sup>3</sup>

<sup>1</sup>Smith College

<sup>2</sup>North Carolina State University

<sup>3</sup>Harvard University  
kcook93@smith.edu

The Botswana Combination Prevention Project was a cluster-randomized trial evaluating the impact of combination HIV prevention on the 3-year cumulative incidence of HIV in Botswana. The trial's follow-up period coincided with Botswana's national adoption of a universal test-and-treat strategy for HIV management. In this talk, we set out to determine whether, and to what extent, this change in policy (i) modified the observed preventative effects of the study intervention and (ii) was associated with a reduction in the incidence of HIV in Botswana. To address these questions, we propose a marginal proportional hazards model for clustered interval-censored data with time-dependent covariates. We develop a composite expectation maximization algorithm that facilitates estimation of the model parameters without placing parametric assumptions on either the baseline hazard functions or the within-cluster dependence structure. We also develop a robust profile composite likelihood-based sandwich estimator for the variance. We discuss both the theoretical properties of these estimators and their performance in a series of simulation studies. We conclude by applying them to a re-analysis of the Botswana Combination Prevention Project, with the national adoption of a universal test-and-treat strategy now modeled as a time-dependent covariate.

#### **Marginal Structural Models for Network-Level Interventions: The Limiting Variance is Smaller with Estimated Weights**

♦*Judith Lok*<sup>1</sup>, *Ashley Buchanan*<sup>2</sup>, *Luis Iberico*<sup>1</sup> and *Donna Spiegelman*<sup>3</sup>

<sup>1</sup>Boston University

<sup>2</sup>University of Rhode Island

<sup>3</sup>Yale University  
jjlok@bu.edu

Network-level interventions/clustering are increasingly common in public health research. In observational studies, where the interventions are not randomized, standard methods to analyze how interventions affect outcomes may not lead to consistent estimators. Marginal Structural Models are often used to estimate effects of time-dependent interventions from observational data. It is easy to show that under no unmeasured confounding, Marginal Structural Models combined with a fixed working covariance matrix lead to consistent, asymptotically normal intervention effect estimates. Confidence interval/variance estimation however is complicated, because the weights underlying Marginal Structural Model

fit are estimated, affecting the variance. Thus, analysts often use the bootstrap, which for larger datasets uses considerable computing time. Rotnitzky and Robins (1995) showed that for independent units/participants subject to censoring, Inverse Probability of Censoring Weighting ignoring estimation of the weights leads to conservative inference. The same has been shown for independent units/participants and point treatments, and for independent participants and time-dependent treatments. In this presentation we show this result for network-level interventions/clustered data and time-dependent treatments. We also provide a correction to the conservative inference. We illustrate our methods with HPTN 037, a network-level study for how repeated peer education affects HIV risk behavior over time in Philadelphia.

#### **Model-Robust and Efficient Covariate Adjustment for Cluster-Randomized Experiments**

*Bingkai Wang*<sup>1</sup>, *Chan Park*<sup>1</sup>, *Dylan Small*<sup>1</sup> and ♦*Fan Li*<sup>2</sup>

<sup>1</sup>The Statistics and Data Science Department of the Wharton School, University of Pennsylvania, Philadelphia, PA, USA

<sup>2</sup>Department of Biostatistics, Yale University School of Public Health, New Haven, CT, USA  
fan.f.li@yale.edu

Cluster-randomized experiments are increasingly used to evaluate interventions in routine practice conditions, and researchers often adopt model-based methods with covariate adjustment in the statistical analyses. However, the validity of model-based covariate adjustment is unclear when the working models are misspecified, leading to ambiguity of estimands and risk of bias. In this presentation, we first adapt two conventional model-based methods, generalized estimating equations and linear mixed models, with weighted g-computation to achieve robust inference for cluster-average and individual-average treatment effects. Furthermore, we propose an efficient estimator for each estimand that allows for flexible covariate adjustment and additionally addresses cluster size variation dependent on treatment assignment and other cluster characteristics. Such cluster size variations often occur post-randomization and, if ignored, can lead to bias of model-based estimators. For our proposed estimator, we prove that when the nuisance functions are consistently estimated by machine learning algorithms, the estimator is consistent, asymptotically normal, and efficient. When the nuisance functions are estimated via parametric working models, the estimator is triply-robust. Simulation studies and analyses of three real-world cluster-randomized experiments demonstrate that the proposed methods are superior to existing alternatives.

#### **Session 23CHI91: Recent Development in Estimating Treatment Effects with Complex Medical Data**

##### **Examining Subgroup-Specific Treatment Effects in Multi-Source Data: Source-Specific Inference and Transportability to an External Population**

♦*Guanbo Wang*<sup>1</sup>, *Alex Levis*<sup>2</sup> and *Issa Dahabreh*<sup>1</sup>

<sup>1</sup>Harvard University

<sup>2</sup>Carnegie Mellon University  
awang.08040222@gmail.com

One major challenge in the estimation of effect heterogeneity is that the sample size of the data used is typically not enough to precisely capture how effects vary according to the effect modifiers. Therefore, there is interest in synthesizing evidence across multi-source data (e.g., multi-center trials, meta-analyses of randomized trials, pooled analyses of observational cohorts) to improve the precision

## Abstracts

---

of estimators of heterogeneous treatment efficacy. Furthermore, when combining information from multi-source data, the samples typically do not represent a common target population of substantive interest. This raises the question of how to combine information from multi-source data in a way that is interpretable in the context of some meaningful target population of interest while using evidence across multi-source data to improve efficiency. We develop and evaluate methods for using multi-source data to estimate subgroup treatment effects in an external target population or the populations underlying the data sources. We propose a doubly robust estimator that under mild conditions is non-parametrically efficient and allows for nuisance functions to be estimated using machine learning methods. We illustrate the methods in meta-analyses of randomized trials for schizophrenia and bipolar disorder.

### **Modeling Interval-Censored Outcome Data with an Interval-Censored Covariate with Application to Hiv Viral Bound Research**

♦Dongdong Li<sup>1</sup>, Yue Song<sup>2</sup>, Wenbin Lu<sup>3</sup>, Huldrych Gunthard<sup>4</sup>, Roger Kouyoumji<sup>4</sup> and Rui Wang<sup>5</sup>

<sup>1</sup>Harvard Medical School

<sup>2</sup>Harvard T. H. Chan School of Public Health

<sup>3</sup>North Carolina State University

<sup>4</sup>University of Zurich

<sup>5</sup>Harvard Medical School and Harvard T. H. Chan School of Public Health

dongdong\_li@hphci.harvard.edu

Motivated by the need to assess whether the time to viral suppression after antiretroviral therapy (ART) initiation is predictive of the time to viral rebound after ART interruption, we investigate modeling approaches relating an interval-censored outcome and an interval-censored covariate. We develop estimation and inference procedures for fitting a proportional hazards regression model when both the outcome and a covariate can be interval-censored, without making parametric distributional assumptions about baseline hazard functions, through the use of an Expectation-Maximization algorithm. We evaluate the finite-sample performance of the proposed method through simulation studies. To illustrate, we assess the effect of time to viral suppression after ART initiation on time to viral rebound after ART interruption using data from the Zurich Primary HIV Infection cohort.

### **Causal Inference for Assessing Cost-Effectiveness with Multi-State Modelling**

♦Yi Xiong<sup>1</sup>, Gary Chan<sup>2</sup> and Li Hsu<sup>3</sup>

<sup>1</sup>University of Manitoba; Fred Hutchinson Cancer Center

<sup>2</sup>University of Washington

<sup>3</sup>Fred Hutchinson Cancer Center  
yi.xiong@umanitoba.ca

Health-care policy makers are often interested in the cost-effectiveness of an intervention. The effectiveness is usually measured by quality adjusted life years, which is subject to informative censoring, and the costs, both of which are often assessed from large-scale observational studies and databases (e.g., claims data, large cohort studies) and are thus susceptible to confounding. There is considerably rich literature available to accommodate censoring and adjust for confounding factors. However, most cost-effectiveness studies are primarily concerned with the terminal event rather than the entire disease progression. This paper is motivated by informing optimal initial screening age for colorectal cancer (CRC) through cost-effectiveness analysis. We provide a unified measure of cost-effectiveness with semi-competing risks

and multistate modeling, which allows us to gain insights on benefit and cost at each stage of cancer progression. Unlike most existing causal inference works focusing on static interventions, we develop a causal framework to evaluate cost-effectiveness as a function of time-varying screening strategy. These methods are justified theoretically and numerically using both simulation and the CRC data from the Women's Health Initiative observational study. This is a joint work with Dr. Li Hsu (Fred Hutchinson Cancer Center) and Dr. Gary Chan (University of Washington).

### **Session 23CHI92: Advanced Methods for Analyzing Categorical Data**

#### **Inference for Seemingly Unrelated Linear Mixed Models**

♦Lichun Wang and Yang Yang

Department of Statistics, Beijing Jiaotong University  
wlc@amss.ac.cn

Linear mixed models and data generated from repeated measurements have found substantial applications in many disciplines. This paper considers the estimation problem of fixed effects and variance components in the system of two seemingly unrelated linear mixed models. We first propose the covariance adjustment estimator of the fixed effects and then construct consistent estimators of its unknown parameters and further study the small sample performance of the two-stage estimator. It is shown that the proposed two-stage covariance adjustment estimator is preferred and feasible based on its theoretical properties and some numerical illustrations. Finally, we investigate the effects of covariance adjustment on the variance components by plugging the two-stage estimator into the restricted log-likelihood of the variance components, and the results imply that seemingly unrelated linear mixed models may only improve the estimation of the fixed effects.

#### **Classification with Imbalanced Data in Medicine**

Wanhua Su

MacEwan University  
suw3@macewan.ca

Detecting rare but critical healthcare events in massive data is a common task in biostatistics and medicine. This can be treated as a classification problem with imbalanced outcomes in which the prevalence of the event is extremely low. There are two popular approaches to handling classification with imbalanced data: using cost-sensitive performance metrics to train the classifiers and resampling methods such as SMOTE (synthetic minority oversampling technique). We proposed a cluster-based under-sampling method to address the classification problem with imbalanced data. The results based on simulation studies and a variety of benchmark data sets show that the proposed method outperforms the SMOTE in both areas under the ROC and Precision-Recall curves.

#### **Iteratively Reweighted Least Squares Method for Estimating Polyserial and Polychoric Correlation Coefficients**

Peng Zhang<sup>1</sup>, ♦Ben Liu<sup>1</sup> and Jingjing Pan<sup>2</sup>

<sup>1</sup>School of Mathematical Sciences, Zhejiang University

<sup>2</sup>Zhejiang Super Soul Artificial Intelligence Research Institute  
12035024@zju.edu.cn

An iteratively reweighted least squares (IRLS) method is proposed for the estimation of polyserial and polychoric correlation coefficients in this paper. It calculates the slopes in a series of weighted linear regression models fitting on conditional expected values. For polyserial correlation, conditional expectations of the latent predictor is derived from the observed ordinal categorical variable, and the

regression coefficient is obtained with the weighted least squares method. In estimating polychoric correlation coefficient, conditional expectations of the response variable and the predictor are updated in turns. Standard errors of the estimators are obtained using the delta method based on data summaries instead of the whole data. Conditional univariate normal distribution is exploited and a single integral is numerically evaluated in the proposed algorithm, comparing to the double integral computed numerically based on the bivariate normal distribution in the traditional maximum likelihood (ML) approaches. This renders the new algorithm very fast in the estimation of both polyserial and polychoric correlation coefficients. Thorough simulation studies are conducted to compare the performances of the proposed method with the classical ML methods. Real data analyses illustrate the advantage of the new method in computing speed.

### **Modeling Clustered Binary Data with an Application to Stock Crash Analysis**

*Ruixi Zhao<sup>1</sup>, Renjun Ma<sup>1</sup>, ♦Guohua Yan<sup>1</sup>, Haomiao Niu<sup>2</sup> and Wenjiang Jiang<sup>3</sup>*

<sup>1</sup>University of New Brunswick

<sup>2</sup>University of Siena

<sup>3</sup>Yunan Normal University

gyan@unb.ca

Various random effects models have been developed for clustered binary data; however, traditional approaches to these models generally rely heavily on the specification of a continuous random effects distribution such as Gaussian or beta distribution. In this work, we introduce a new model that incorporates nonparametric unobserved random effects on unit interval (0,1) into logistic regression multiplicatively with fixed effects. This multiplicative model setup facilitates the prediction of nonparametric random effects and corresponding model interpretations. A distinctive feature of our approach is that a closed-form expression has been derived for the predictor of nonparametric random effects on unit interval (0,1) in terms of known covariates and responses. A quasi-likelihood approach has been developed in the estimation of the model. The results are robust against random effects distributions from very discrete binary to continuous beta distributions. We illustrate our method by analyzing recent large stock crash data in China. The performance of our method is also evaluated through simulation studies.

### **Session 23CHI93: Limit Theorems and Inference for Time Series and Spatial Processes**

#### **Limit Theory for Autoregressive Processes with Roots Close to Unity**

*Nannan Ma<sup>1</sup>, Hailin Sang<sup>2</sup> and ♦Guangyu Yang<sup>3</sup>*

<sup>1</sup>Agricultural Bank of China

<sup>2</sup>University of Mississippi

<sup>3</sup>Zhengzhou University  
guangyu@zzu.edu.cn

We establish the asymptotic theory of least absolute deviation estimators for AR(1) processes with autoregressive parameter satisfying  $n(\rho_n - 1) \rightarrow \gamma$  for some fixed  $\gamma$  as  $n \rightarrow \infty$ , which is parallel to the results of ordinary least squares estimators developed by Andrews and Guggenberger (*Journal of Time Series Analysis*, **29**, 203-212, 2008) in the case  $\gamma = 0$  or Chan and Wei (*Annals of Statistics*, **15**, 1050-1063, 1987) and Phillips (*Biometrika*, **74**, 535-574, 1987) in the case  $\gamma \neq 0$ . Simulation experiments are conducted to confirm

the theoretical results and to demonstrate the robustness of the least absolute deviation estimation.

#### **Local Limit Theorem for Linear Random Fields**

*Timothy Fortune<sup>1</sup>, Magda Peligrad<sup>2</sup>, ♦Hailin Sang<sup>3</sup>, Yimin Xiao<sup>4</sup> and Guangyu Yang<sup>5</sup>*

<sup>1</sup>Nicholls State University

<sup>2</sup>University of Cincinnati

<sup>3</sup>University of Mississippi

<sup>4</sup>Michigan State University

<sup>5</sup>Zhengzhou University

sang@olemiss.edu

We establish local limit theorems for linear random fields when the i.i.d. innovations have finite second moment or the innovations have infinite second moment and belong to the domain of attraction of a stable law with index  $0 < \alpha \leq 2$  under the condition that the innovations are centered if  $1 < \alpha \leq 2$  and are symmetric if  $\alpha = 1$ . When the coefficients are absolutely summable we do not have restriction on the regions of summation. However, when the coefficients are not absolutely summable we add the variables on unions of rectangles and we impose regularity conditions on the coefficients depending on the number of rectangles considered. Our results are new also for the dimension 1, i.e. for linear sequences of random variables. The examples include the fractionally integrated processes for which the results of a simulation study is also included. This talk is based on two papers jointly with Timothy Fortune, Magda Peligrad, Yimin Xiao and Guangyu Yang.

#### **Kernel Entropy Estimation for Long Memory Linear Processes with Infinite Variance**

*Hui Liu and ♦Fangjun Xu*

East China Normal University

fjxu@finance.ecnu.edu.cn

Let  $X = \{X_n : n \in \mathbb{N}\}$  be a long memory linear process with innovations in the domain of attraction of an  $\alpha$ -stable law ( $0 < \alpha < 2$ ). Assume that the linear process  $X$  has a bounded probability density function  $f(x)$ . Then, under certain conditions, we consider the estimation of the quadratic functional  $\int_{\mathbb{R}} f^2(x) dx$  by using the kernel estimator

$$T_n(h_n) = \frac{2}{n(n-1)h_n} \sum_{1 \leq j < i \leq n} K\left(\frac{X_i - X_j}{h_n}\right).$$

The simulation study for long memory linear processes with symmetric  $\alpha$ -stable innovations is also given.

#### **Self-Normalized Cramer Type Moderate Deviations for Martingales with Applications**

*♦Xiequan Fan<sup>1</sup> and Qiman Shao<sup>2</sup>*

<sup>1</sup>Northeastern University at Qinhuangdao

<sup>2</sup>Southern University of Science and Technology  
fanxiequan@hotmail.com

Cramer type moderate deviations give a quantitative estimate for the relative error of the normal approximation and provide theoretical justifications for many estimator used in statistics. In this talk, we talk about self-normalized Cramer type moderate deviations for martingales under some mild conditions. Applications of our result to Student's statistic, stationary martingale difference sequences and branching processes in a random environment are discussed.

## Abstracts

---

### Session 23CHI96: New Advances in Causal Inference and Missing Data

#### Bayesian Machine Learning g-Computation Formula for Survival Data

♦*Xinyuan Chen<sup>1</sup>, Liangyuan Hu<sup>2</sup> and Fan Li<sup>3</sup>*

<sup>1</sup>Mississippi State University

<sup>2</sup>Rutgers University

<sup>3</sup>Yale University

xchen@math.msstate.edu

We develop a Bayesian machine learning approach for longitudinal predictive and causal inference via the *g*-computation formula. We use the Bayesian additive regression trees (BART) for the model specification of the time-evolving generative components to reduce bias caused by parametric modelling. Theoretically, we introduce new *g*-computation formulas leveraging the concept of the longitudinal balancing score (BS), which essentially is an effective dimension reduction tool when the number of confounders increases. The formulation of the longitudinal BS is adaptive according to different types of treatment regimes, i.e., deterministic or dynamic. Simulation studies were conducted to evaluate the performance of the proposed method in various scenarios. Our work is motivated by the Multi-Ethnic Study of Atherosclerosis (MESA) from the National Heart, Lung and Blood Institute (NHLBI).

#### Estimation of Causal Influence Effect in Social Networks with Latent Location Adjustment

♦*Samrachana Adhikari and Seungha Um*

NYU School of Medicine

samrachana.adhikari@nyulangone.org

Similarities in health behavior, disease state, and outcomes of two individuals in a network relationship can be caused by three primary mechanisms: contagion or influence, homophilous selection, and common social or environmental factors. However, it is often challenging to separate the effect of influence from other processes that operate simultaneously, including homophily, especially when factors leading to homophily are unobserved. To estimate latent homophily adjusted causal influence effect we present a two-stage modeling framework that leverages estimates of latent locations from a class of conditionally independent network models.

#### Estimating the Causal Effect of a Longitudinal Treatment when Covariates are Subject to Missing Data

♦*Liangyuan Hu<sup>1</sup> and Jungang Zou<sup>2</sup>*

<sup>1</sup>Rutgers University

<sup>2</sup>Columbia University

liangyuan.hu@rutgers.edu

Missing data are a pervasive problem in complex longitudinal datasets, and pose a substantial challenge for drawing causal inferences about longitudinal treatments. For missing at random longitudinal covariates, existing imputation methods are primarily based on parametric models, in which exact relationships among longitudinal response, treatment, and covariates are made explicit. Incorrect specification of the parametric form can introduce model misspecification biases. We propose flexible semi- and non-parametric Bayesian sequential imputation methods for missing at random longitudinal covariates. We first develop a novel Bayesian trees mixed-effects model to flexibly model the longitudinal trajectories. We then develop an efficient MCMC algorithm to sequentially impute the missing longitudinal covariates data using the developed model. The novel longitudinal missing data methodology is then formally integrated with *g*-computation to study the causal effect of longitudinal treatment. We conduct expansive simulations to investigate

the practical operating characteristics of our proposed methods. Finally we apply our methods to a NHLBI study dataset to estimate and validate optimal dynamic antihypertensive treatment initiating rules.

### Session 23CHI98: Analysis of Complex Network and Text Data

#### Bayesian Inference of Spatially Varying Correlations via the Thresholded Correlation Gaussian Process

♦*Moyan Li<sup>1</sup>, Jian Kang<sup>1</sup> and Lexin Li<sup>2</sup>*

<sup>1</sup>University of Michigan

<sup>2</sup>University of Berkley

moyanli@umich.edu

A central question in multimodal neuroimaging analysis is to understand the association between two imaging modalities and to identify brain regions where such an association is statistically significant. In this article, we propose a Bayesian nonparametric spatially varying correlation model to address estimation and inference of such regions. We build our model based on the thresholded correlation Gaussian process, which ensures piecewise smoothness, sparsity, and jump discontinuity of spatially varying correlations, and works well even when the number of subjects is limited or the signal-to-noise ratio is small. We study the identifiability of our model, establish the large support property, and derive the posterior consistency and selection consistency. We also develop a highly efficient Gibbs sampler and its variant to compute the posterior distribution. We illustrate the method with both simulations and an analysis of functional magnetic resonance imaging data from the Human Connectome Project.

#### Stock Co-Jump Networks

♦*Yi Ding<sup>1</sup>, Guoli Li<sup>2</sup>, Yingying Li<sup>2</sup> and Xinghua Zheng<sup>2</sup>*

<sup>1</sup>University of Macau

<sup>2</sup>Hong Kong University of Science and Technology

ydingust@gmail.com

We propose a Degree-Corrected Block Model with Dependent Multivariate Poisson edges (DCBM-DMP) to study stock co-jump dependence. To estimate the community structure, we extend the SCORE algorithm in Jin (2015) and develop a Spectral Clustering On Ratios-of-Eigenvectors for networks with Dependent Multivariate Poisson edges (SCORE-DMP) algorithm. We prove that SCORE-DMP enjoys strong consistency in community detection. Empirically, using high-frequency data of S & P 500 constituents, we construct two co-jump networks according to whether the market jumps and find that they exhibit different community features than GICS. We further show that the co-jump networks help in stock return prediction.

### Session 23CHI99: Advances in Bayesian Adaptive Design Methods for Drug Development

#### From Finding Maximum Tolerated Dose (Mtd) to Determining Optimal Biological Dose (Obd) in Oncology Drug Development

*J. Jack Lee*

University of Texas MD Anderson Cancer Center

j.lee@mdanderson.org

Traditional oncology drug development program is based on cytotoxic chemotherapies. The underlying assumption is that the higher the dose, the more efficacious the drug is. The paradigm is to find the maximum tolerate dose (MTD) or the recommended Phase 2

dose (RP2D) in Phase I studies followed by testing the drug efficacy at MTD or RP2D in Phase II studies. This paradigm, however, may not be suitable for molecularly targeted agents and immunotherapies because higher dose can lead to higher toxicity but may not result in higher efficacy. In 2022, US FDA launch “Project Optimus” to reform the dose optimization and dose selection paradigm in oncology. In this talk, I will introduce the model-assisted designs, which incorporate Bayesian modeling with good statistical operating characteristics, yet easy to implement by listing precalculated boundaries to find the optimal biological dose (OBD). Three designs based on the Bayesian Optimal INterval (BOIN) design: the one-stage BOIN12 design, the two-stage utility-based U-BOIN design, and the TITE-BOIN12 design for late-onset toxicity and efficacy will be discussed. These three designs are effective in finding the OBD. Examples will be given. User-friendly free software will be demonstrated.

#### **Demo: Bayesian Adaptive Dose Exploration–monitoring–optimization Design Based on Short, Intermediate, and Long-Term Outcomes**

Ruitao Lin

The University of Texas MD Anderson Cancer Center  
rlin@mdanderson.org

The US Food and Drug Administration (FDA) launched Project Optimus and released the draft guideline to reform the dose selection and optimization process. The gold-standard endpoint to evaluate a cancer treatment is a long-term survival endpoint, such as the overall survival or progression-free survival. An optimal dose recommended for late phases should ultimately possess a promising enough survival profile. Existing early-phase trial designs that rely on short-term toxicity and efficacy data may identify a dose that may result in suboptimal long-term survival benefits. However, using long-term survival endpoint in trial monitoring requires a longer follow-up time, and thus results in a prolonged trial duration. To address these challenges, a generalized dose optimization procedure consisting of three seamlessly connected stages is proposed. In the first stage, the short-term pharmacodynamics biomarker and toxicity endpoints are utilized to screen out overly toxic doses and doses that do not yield bioactivity. In the second stage, patients are treated at admissible doses identified in the first stage to collect accumulating toxicity and efficacy data for dose monitoring. In the third stage, patients are randomized to the doses in the refined admissible dose set to identify the optimal dose through restricted mean survival time.

#### **A Bayesian Adaptive Phase i/Ii Clinical Trial Design with Late-Onset Competing Risk Outcomes**

♦Yifei Zhang<sup>1</sup>, Sha Cao<sup>2</sup>, Chi Zhang<sup>2</sup>, Ick Hoon Jin<sup>3</sup> and Yong Zang<sup>2</sup>

<sup>1</sup>Jiangsu Hengrui Pharmaceuticals Co., Ltd.

<sup>2</sup>Indiana University,Biostatistics & Health Data Science

<sup>3</sup>Yonsei University  
yifei.zhang@hengrui.com

Early-phase dose-finding clinical trials are often subject to the issue of late-onset outcomes. In phase I/II clinical trials, the issue becomes more intractable because toxicity and efficacy can be competing risk outcomes such that the occurrence of the first outcome will terminate the other one. In this paper, we propose a novel Bayesian adaptive phase I/II clinical trial design to address the issue of late-onset competing risk outcomes. We use the continuation-ratio model to characterize the trinomial response outcomes and the cause-specific hazard rate method to model the competing-risk sur-

vival outcomes. We treat the late-onset outcomes as missing data and develop a Bayesian data augmentation method to impute the missing data from the observations. We also propose an adaptive dose-finding algorithm to allocate patients and identify the optimal biological dose during the trial. Simulation studies show that the proposed design yields desirable operating characteristics.

#### **Shotgun-2: a Bayesian Phase i/Ii Basket Trial Design to Identify Indication-Specific Optimal Biological Doses**

Fangrong Yan  
13149306279@163.com

For novel molecularly targeted agents and immunotherapies, the objective of dose-finding is often to identify the optimal biological dose, rather than the maximum tolerated dose. However, optimal biological doses may not be the same for different indications, challenging the traditional dose-finding framework. Therefore, we proposed a Bayesian phase I/II basket trial design, named “shotgun-2,” to identify indication-specific optimal biological doses. A dose-escalation part is conducted in stage I to identify the maximum tolerated dose and admissible dose sets. In stage II, dose optimization is performed incorporating both toxicity and efficacy for each indication. Simulation studies under both fixed and random scenarios show that, compared with the traditional “phase I + cohort expansion” design, the shotgun-2 design is robust and can improve the probability of correctly selecting the optimal biological doses. Furthermore, this study provides a useful tool for identifying indication-specific optimal biological doses and accelerating drug development.

#### **Session 23CHIKT2: Keynote Lecture 2**

##### **Theory of FPCA for Discretized Functional Data**

Fang Yao  
Peking University

Functional data analysis is an important research field in statistics which treats data as random functions drawn from some infinite-dimensional functional space, and functional principal component analysis (FPCA) plays a central role for data reduction and representation. After nearly three decades of research, there remains a key problem unsolved, namely, the perturbation analysis of covariance operator for diverging number of eigencomponents obtained from noisy and discretely observed data. This is fundamental for studying models and methods based on FPCA, while there has not been much progress since the result obtained by Hall et al. (2006) for a fixed number of eigenfunction estimates. In this work, we establish a unified theory for this problem, deriving the moment bounds of eigenfunctions and asymptotic distributions of eigenvalues for a wide range of sampling schemes. We also exploit double truncation to derive the uniform convergence of such estimated eigenfunctions. The technical arguments in this work are useful for handling the perturbation series of discretely observed functional data and can be applied in models and methods involving inverse using FPCA as regularization, such as functional linear regression.

#### **Session 23CHI123: Recent Developments in Causal Inference**

##### **Estimation and Inference of a Directed Acyclic Graph using Gwas Summary Statistics**

Rachel Zilinskas<sup>1</sup>, ♦Chunlin Li<sup>2</sup>, Xiaotong Shen<sup>2</sup>, Wei Pan<sup>3</sup> and

## Abstracts

---

Tianzhong Yang<sup>3</sup>

<sup>1</sup>Statistics and Data Corporation

<sup>2</sup>School of Statistics, University of Minnesota

<sup>3</sup>Division of Biostatistics, University of Minnesota  
li000007@umn.edu

Estimating phenotype networks is a growing field in computational biology. It deepens the understanding of disease etiology and is useful in many applications. In this talk, I will present a method that constructs a phenotype network by assuming a Gaussian linear structure model embedding a directed acyclic graph (DAG). We utilize genetic variants as instrumental variables and show how our method only requires access to summary statistics from a genome-wide association study (GWAS) and a reference panel of genotype data. Besides estimation, a distinct feature of the method is its summary statistics-based likelihood ratio test on directed edges. We applied our method to estimate a directed network of 29 cardiovascular-related proteins and linked the estimated network to Alzheimer's disease (AD).

### Local Method for Causal Effect Estimation

Yue Liu

liuyue\_stats@ruc.edu.cn

Evaluation of causal effects, including total effect, direct effect, and indirect effect, is critical in comparative effectiveness research on clinical interventions. However, it is well established that identifying causal effects from observational data is challenging when a causal directed acyclic graph is absent. To overcome this limitation and leverage the potential of large-scale and comprehensive observational healthcare data, we propose methodologies that combine causal structure learning methods with causal effect estimation algorithms. Our approaches aim to identify all possible causal effects from both the estimated graphs (MPDAG and CPDAG) and their local structures. Furthermore, we extend our approaches to address algorithmic fairness in machine learning. The experimental results show that, our algorithms possible to estimate all possible causal effects efficiently.

### Learning Linear Non-Gaussian Directed Acyclic Graph with Diverging Number of Nodes

Ruixuan Zhao, ♦Xin He and Junhui Wang

he.xin17@mail.shufe.edu.cn

An acyclic model, often depicted as a directed acyclic graph (DAG), has been widely employed to represent directional causal relations among collected nodes. In this article, we propose an efficient method to learn linear non-Gaussian DAG in high dimensional cases, where the noises can be of any continuous non-Gaussian distribution. The proposed method leverages the concept of topological layer to facilitate the DAG learning, and its theoretical justification in terms of exact DAG recovery is also established under mild conditions. Particularly, we show that the topological layers can be exactly reconstructed in a bottom-up fashion, and the parent-child relations among nodes can also be consistently established. The established asymptotic DAG recovery is in sharp contrast to that of many existing learning methods assuming parental faithfulness or ordered noise variances. The advantage of the proposed method is also supported by the numerical comparison against some popular competitors in various simulated examples as well as a real application on the global spread of COVID-19.

### Introducing the Specificity Score: a Measure of Causality Beyond p Value

Wang Miao

Peking University  
mwfy@pku.edu.cn

There is considerable debate and doubt about the use of P value in scientific research in recent years, particularly after its use is banned in several prestigious journals. Much scientific research is concerned with uncovering causal associations, however, P value by definition is a measure of the significance of a statistical association, which could be biased from the causal association of interest and lead to false discoveries due to confounding. In this talk, I will introduce a score measuring the specificity of causal associations and a specificity score-based test about the existence of causal effects, in the presence of unmeasured confounding. Under certain conditions, this approach has controlled type I error and power approaching unity for testing the null hypothesis of no causal effect. This approach is particularly suitable for joint causal discovery with multiple treatments and multiple outcomes, such as gene expressions studies, Mendelian randomization and EHR studies. A visualization approach using a specificity map is proposed to communicate all specificity score/test information in a universal and effective manner. Identification and estimation will be briefly covered. Simulations are used for illustration and an application to a mouse obesity dataset detects potential active effects of genes on clinical traits that

## Session 23CHI130: Space-Filling Designs (I)

### Column Expanded Latin Hypercube Designs

♦Qiao Wei, Jian-Feng Yang and Min-Qian Liu

Nankai University  
785135798@qq.com

Maximin distance designs and orthogonal designs are extensively applied in computer experiments, but the construction of such designs is challenging, especially under the maximin distance criterion. In this paper, by adding columns to a fold-over optimal maximin L<sub>2</sub>-distance Latin hypercube design (LHD), we construct a class of LHDs, column expanded Latin hypercube designs, which are nearly optimal under both the maximin L<sub>2</sub>-distance and orthogonality criteria. The advantage of the proposed method is that the resulting designs can have flexible numbers of factors without computer search. Detailed comparisons with existing LHDs show that the constructed LHDs have larger minimum distances between design points and smaller correlation coefficients between distinct columns.

### Construction of Orthogonal Doubly Coupled Designs

Mengmeng Liu, ♦Jinyu Yang and Min-Qian Liu

Nankai University  
jyyang@nankai.edu.cn

Computer experiments with both quantitative and qualitative factors occur in many scientific and engineering applications. How to choose suitable designs for such experiments is an important issue. Sliced Latin hypercube designs (LHDs) are the first proposed designs for this issue. The increase of level combinations of the qualitative factors will lead to a large increasing cost of the sliced LHD. For the reason of run size economy, marginally coupled designs were proposed without the ability to capture the effects between any two (and more) qualitative factors and quantitative factors. Recently, to balance the run sizes and design efficiencies to estimate the interactions between any two qualitative factors and the quantitative factors, doubly coupled designs (DCDs) were introduced by Yang et al. (2022). Orthogonality is an important property for the design of computer experiment. In this paper, we first propose methods to construct orthogonal DCDs (ODCDs) with s<sup>t</sup> (t = 3) runs, where s > 2 is a prime or prime power. Furthermore, we introduce two constructions of DCDs with desirable space-filling

properties. Most of the constructed designs can accommodate the maximum number of qualitative factors and a relatively large number of quantitative factors.

#### A-Optimal Designs for Non-Parametric Symmetrical Global Sensitivity Analysis

**♦Xueping Chen<sup>1</sup>, Yujie Gai<sup>2</sup> and Xiaodi Wang<sup>2</sup>**

<sup>1</sup>Jiangsu University of Technology

<sup>2</sup>Central University of Finance and Economics  
chenxueping@jsut.edu.cn

In the early stage of exploring a complex system, a preliminary experiment is used to capture the key characteristics of the model. Symmetrical global sensitivity analysis (SGSA) is one such experiment that explores the symmetrical structure of model by decomposing the model into independent symmetric functions. However, the existing experimental plans for SGSA rely on deterministic computational models that produce unique values of outputs when executed for specific values of inputs. In this paper, the problem of designing experiments for non-parametric SGSA is considered. Here the phrase “non-parametric” refers to model outputs containing random errors. The main result in the paper shows that a symmetrical design with certain constraints achieves A-optimum for the estimation of each output element function, and guarantees the superiority of the SGSA result. The statistical properties of non-parametric SGSA based on the optimal designs are further discussed, showing that the non-influential sensitivity indices can be estimated with low bias and volatility. Two explicit structures of the optimal designs are obtained. The optimality of the derived design is validated by simulation in the end.

#### Construction of a Class of Nested Orthogonal Arrays

*Shanqi Pang*

Henan Normal University  
pangshanqi@263.net

TBC

#### Session 23CHI135: Tree and Graphical Models

##### Chain Graph Models: Identifiability, Estimation and Asymptotics

*Ruixuan Zhao<sup>1</sup>, Haoran Zhang<sup>2</sup> and ♦Junhui Wang<sup>2</sup>*

<sup>1</sup>City University of Hong Kong

<sup>2</sup>Chinese University of Hong Kong  
junhuiwang@cuhk.edu.hk

In this talk, we consider a flexible chain graph (CG) model, which admits both undirected and directed edges in one graph and thus can encode much more diverse relations among objects. We first establish the identifiability conditions for the CG model through a low rank plus sparse matrix decomposition, where the sparse matrix implies the sparse undirected edges within each chain component and the low rank matrix implies the presence of hub nodes with multiple children or parents. On this ground, we develop an efficient estimation method for reconstructing the CG structure, which first identifies the chain components via estimated undirected edges, determines the causal ordering of the chain components, and eventually estimates the directed edges among the chain components. Its theoretical properties will be discussed in terms of both asymptotic and finite-sample probability bounds on model estimation and graph reconstruction. The advantage of the proposed method is also demonstrated through extensive numerical experiments on both synthetic data and the Standard & Poor's 500 index data.

##### Confidence Band Estimation of Random Survival Forests

*Sarah Formentini<sup>1</sup>, Wei Liang<sup>2</sup> and ♦Ruqing Zhu<sup>1</sup>*

<sup>1</sup>University of Illinois Urbana Champaign

<sup>2</sup>Xiamen University  
rqzhu@illinois.edu

Survival random forest is a popular machine learning tool for modeling censored survival data. However, there is currently no statistically valid and computationally feasible approach for estimating its confidence band. This paper proposes an unbiased confidence band estimation by extending recent developments in infinite-order incomplete U-statistics. The idea is to estimate the variance-covariance matrix of the cumulative hazard function prediction on a grid of time points. We then generate the confidence band by viewing the cumulative hazard function estimation as a Gaussian process whose distribution can be approximated through simulation. This approach is computationally easy to implement when the subsampling size of a tree is no larger than half of the total training sample size. Numerical studies show that our proposed method accurately estimates the confidence band and achieves desired coverage rate. We apply this method to veterans' administration lung cancer data.

##### Joint Modeling of Change-Point Identification and Dependent Dynamic Community Detection

**♦Diqing Li<sup>1</sup>, Yubai Yuan<sup>2</sup>, Xinsheng Zhang<sup>3</sup> and Annie Qu<sup>4</sup>**

<sup>1</sup>Zhejiang Gongshang University

<sup>2</sup>the Pennsylvania State University

<sup>3</sup>Fudan University

<sup>4</sup>University of California Irvine  
dqli@mail.zjgsu.edu.cn

The field of dynamic network analysis has recently seen a surge of interest in community detection and evolution. However, existing methods for dynamic community detection do not consider dependencies between edges, which could lead to a loss of information when detecting community structures. In this study, we investigate the problem of identifying a change-point with abrupt changes in the community structure of a network. To do so, we propose an approximate likelihood approach for the change-point estimator and for identifying node membership that integrates marginal information and dependencies of network connectivities. We propose an expectation-maximization-type algorithm that maximizes the approximate likelihood jointly over change-point and community membership evolution. From a theoretical viewpoint, we establish estimation consistency under the regularity condition, and show that the proposed estimators achieve a higher convergence rate than those of their marginal likelihood counterparts, which do not incorporate dependencies between edges. We demonstrate the validity of the proposed method by applying it to the ADHD-200 data set to detect brain functional community changes over time.

##### Tree Building via Hypothesis Tests

*Ze Gao<sup>1</sup>, Bo Zhang<sup>2</sup>, Jiaqi Hu<sup>2</sup>, Tingyin Wang<sup>2</sup>, Heping Zhang<sup>3</sup> and ♦Xueqin Wang<sup>2</sup>*

<sup>1</sup>Sun Yat-sen University

<sup>2</sup>University of Science and Technology of China

<sup>3</sup>Yale University

wangxq20@ustc.edu.cn

Tree building, often viewed as an algorithmic modeling technique, is rapidly expanding and has become a crucial component of machine learning in the last decade. Due to its unique algorithmic characteristics, the study of its statistical properties from the standpoint of statistical hypothesis testing has been limited. In this work, a tree is built by recursively testing whether or not nodes are split. We establish the asymptotic distribution of testing and illustrate how Type

## Abstracts

---

1 and Type 2 errors spread across the tree structure. This framework demonstrates how to conduct hypothesis testing in a recursive structure and permits statistical inferences about the tree's complexity. Moreover, we demonstrate the effectiveness of our strategy through several numerical studies.

### Session 23CHI161: Modelling Unstructured Data: Image, Network, Text and Beyond

#### Dimension Reduction for Covariates in Network Data

♦Junlong Zhao<sup>1</sup>, Xiumin Liu<sup>2</sup>, Hansheng Wang<sup>3</sup> and Chenlei Leng<sup>4</sup>

<sup>1</sup>Beijing Normal University

<sup>2</sup>Beijing Technology and Business University

<sup>3</sup>Peking University

<sup>4</sup>University of Warwick

zhaojunlong928@126.com

A problem of major interest in network data analysis is to explain the strength of connections using context information. To achieve this, we introduce a novel approach, called networksupervised dimension reduction, in which covariates are projected onto low-dimensional spaces to reveal the linkage pattern without assuming a model. We propose a new loss function for estimating the parameters in the resulting linear projection, based on the notion that closer proximity in the low-dimension projection corresponds to stronger connections. Interestingly, the convergence rate of our estimator is found to depend on a network effect factor, which is the smallest number that can partition a graph in a manner similar to the graph colouring problem. Our method has interesting connections to principal component analysis and linear discriminant analysis, which we exploit for clustering and community detection. The proposed approach is further illustrated by numerical experiments and analysis of a pulsar candidates dataset from astronomy.

#### Rankseg: a Consistent Ranking-Based Framework for Segmentation

♦Ben Dai<sup>1</sup> and Chunlin Li<sup>2</sup>

<sup>1</sup>The Chinese University of Hong Kong

<sup>2</sup>The University of Minnesota

bendai@cuhk.edu.hk

Segmentation has emerged as a fundamental field of computer vision and natural language processing, which assigns a label to every pixel/feature to extract regions of interest from an image/text. To evaluate the performance of segmentation, the Dice and IoU metrics are used to measure the degree of overlap between the ground truth and the predicted segmentation. In this paper, we establish a theoretical foundation of segmentation with respect to the Dice/IoU metrics, including the Bayes rule and Dice/IoU-calibration, analogous to classification-calibration or Fisher consistency in classification. We prove that the existing thresholding-based framework with most operating losses are not consistent with respect to the Dice/IoU metrics, and thus may lead to a suboptimal solution. To address this pitfall, we propose a novel consistent ranking-based framework, namely RankDice/RankIoU, inspired by plug-in rules of the Bayes segmentation rule. Three numerical algorithms with GPU parallel execution are developed to implement the proposed framework in large-scale and high-dimensional segmentation. We study statistical properties of the proposed framework. We show it is Dice-/IoU-calibrated, and its excess risk bounds and the rate of convergence are also provided. The numerical effectiveness of RankDice/mRankDice is demonstrated in various simulated exam-

ples and real datasets with state-of-the-art deep learning architectures.

#### Multilayer Random Dot Product Graphs: Nonparametric Estimation and Online Change Point Detection

Fan Wang<sup>1</sup>, Wanshan Li<sup>2</sup>, Oscar Madrid<sup>3</sup>, ♦Yi Yu<sup>1</sup> and Alessandro Rinaldo<sup>2</sup>

<sup>1</sup>University of Warwick

<sup>2</sup>Carnegie Mellon University

<sup>3</sup>UCLA

yi.yu.2@warwick.ac.uk

We start with a multilayer random dot product graph (MRDPG) model, which can be seen as an extension of the random dot product graph model to multilayer networks. The MRDPG model is convenient for incorporating nodes' latent positions when understanding connectivity. Modelling a multilayer network as an MRDPG, we further deploy a tensor-based method and demonstrate its superiority over the state-of-the-art methods. Moving from a single MRDPG to a sequence of MRDPGs, we are concerned with an online change point detection problem in dynamic multilayer random dot product graph (D-MRDPG), especially with random latent positions. At every time point, we observe a realisation from an  $L$ -layered MRDPG. Across layers, we assume that common node sets and random latent positions are shared, but allow for different connectivity matrices. Formulating the realisation as an adjacency tensor, the connectivity is characterised by an  $L$ -dimensional distribution. Across time, we assume that the distribution sequence possesses an abrupt change. Our ultimate goal is to detect this distributional change in an online fashion. With control of false alarms, we aim to detect the change with minimal delay. We devise a novel nonparametric change point detection algorithm, with a kernel estimator in its core.

### Session 23CHI28: Recent Developments on Machine Learning and Precision Medicine

#### Multivariate Change Point Detection for Heterogeneous Series

Yuxuan Guo<sup>1</sup>, Ming Gao<sup>2</sup> and ♦Xiaoling Lu<sup>1</sup>

<sup>1</sup>Renmin University of China

<sup>2</sup>The University of Chicago

xiaolinglu@ruc.edu.cn

The multivariate change point detection problem has been encountered across various fields. Most approaches to this problem assume the series is homogeneous, i.e., all the coordinates change concurrently. Hence, the specific subset of the coordinates containing the change points cannot be determined. In this work, we propose S-MCPD, which is capable of detecting the position of multivariate change points for heterogeneous series by identifying specific coordinates of those changed. Specifically, the problem is discussed in the context of variable selection and transformed into the form of sparse group lasso. In simulation studies, we compared S-MCPD with four existing methods, inspect, sbs, dc, and cpm. The results showed that the performance of S-MCPD was comparable to that of inspect and was superior to other methods in terms of evaluation metrics. In addition, S-MCPD can determine not only the positions of change points, but also the subset of coordinates containing the change points, while other existing methods are unable to achieve this. Moreover, S-MCPD does not depend on the constant variance assumption and works quite well even when the covariance changes, which makes our method more practical. Finally, we applied S-MCPD to two real-world datasets to show its effectiveness.

## **Simultaneous Change Point Detection and Identification for High Dimensional Linear Models**

♦*Bin Liu<sup>1</sup>, Xinsheng Zhang<sup>1</sup> and Yufeng Liu<sup>2</sup>*

<sup>1</sup>Department of Statistics and Data Science, School of Management, Fudan University

<sup>2</sup>Department of Statistics and Operations Research, Department of Genetics and Department of Biostatistics, University of North Carolina at Chapel Hill  
liubin0145@gmail.com

We consider simultaneous change point detection and identification for high dimensional linear models. For change point detection, given any subgroup of variables, we propose a new method for testing the homogeneity of corresponding regression coefficients across the observations. The test statistic is based on a weighted L-infinity norm aggregation, both temporally and spatially, of a de-biased lasso process. A multiplier bootstrap procedure is introduced to approximate its limiting distribution. The proposed bootstrap automatically accounts for the covariance structures of the de-biased lasso process and allows the number of variables to grow exponentially with the sample size. Under some regularity conditions, the proposed new testing procedure controls the type I error asymptotically and is powerful against sparse alternatives and enjoys certain optimality. For change point identification, at each fixed time point, we first aggregate spatial information of the de-biased lasso process with the L-infinity norm, then a change point estimator is obtained by then a change point estimator is obtained by taking “argmax” with respect to time of the above aggregated process. Under H1, the change point estimator is shown to be consistent for the true change point location. Extensive simulation studies justify the validity of the new method.

## **Locally Weighted Nearest Neighbor Classifier**

♦*Guan Yu<sup>1</sup> and Xingye Qiao<sup>2</sup>*

<sup>1</sup>University of Pittsburgh

<sup>2</sup>Binghamton University  
guy24@pitt.edu

Weighted nearest neighbor (WNN) classifiers are fundamental non-parametric classifiers. There exists a vast room of flexibility in the choice of weights for the neighbors in a WNN classifier. In this talk, we will introduce a new locally weighted nearest neighbor (LWNN) classifier which adaptively assigns weights for different test data points. Given a training data set and a test data point  $x_0$ , the weights for classifying  $x_0$  in LWNN are obtained by minimizing an upper bound of the estimation error of the regression function at  $x_0$ . Similar to most other WNN classifiers, LWNN places larger weights on closer neighbors. However, in addition to the ranks of neighbors' distances, the weights in LWNN also depend on the raw values of the distances. Our theoretical study shows that LWNN is minimax optimal when the marginal feature density is bounded away from zero. In the general case with an additional tail assumption on the feature density, the upper bound of the excess risk of LWNN matches the minimax lower bound up to a logarithmic term. Our numerical comparisons between LWNN and some competitors have further demonstrated its effectiveness.

## **Efficient Learning of Optimal Individualized Treatment Rules**

♦*Weibin Mo<sup>1</sup> and Yufeng Liu<sup>2</sup>*

<sup>1</sup>Purdue University

<sup>2</sup>University of North Carolina at Chapel Hill  
harrymok1994@gmail.com

Recent development in data-driven decision science has seen great advancement in individualized decision making. Given the data

with individual covariates, treatment assignments and outcomes, researchers can search for the optimal individualized treatment rule (ITR) that maximizes the expected outcome. Existing methods typically require initial estimation of some nuisance models. The double robustness property that can protect consistency from one nuisance model misspecification out of two is widely advocated. However, when model misspecification exists, especially when the outcome model is misspecified, a doubly robust estimate can be sub-optimal. In this presentation, we discuss such an efficiency loss from a heteroscedasticity perspective. To guarantee optimality under this scenario, we propose an Efficient Learning (E-Learning) framework in the multi-categorical treatment setting. We establish the optimality of E-Learning among the class of doubly robust influence function (IF)-based generalized methods-of-moment (GMM) estimates, which can incorporate most regression-based methods in the ITR literature as special cases.

## **Session 23CHI30: Recent Developments in Biostatistics with their Applications in Cancer Genomics, Screening and Graph Modeling**

### **Bias Correction Models for Ehr Data in the Presence of Non-Random Sampling**

*Judy Zhong*

judy.zhong@nyulangone.org

Electronic health records (EHR) are a source of data including key clinical information with millions of patients, which has been increasingly used for public health research. However, the probability subjects are included in EHR are not random and can depend on various factors, such as demographics, health care referral patterns, and underlying health status. Accordingly, EHR data may not be representative of the target population for inference, which introduces a potential selection bias. Although many studies have shown these issues, little work has been done to develop or apply bias-correction models. In this study, we propose bias correction models for EHR-based research, using aggregated summary data or probability-based survey data of the target population to correct the bias in the association estimates of disease with risk factors and the prevalence of health outcomes. Via simulations under various settings, we demonstrate the efficacy of the proposed method. We apply our method to estimate the prevalence of cardiovascular disease and its associations with risk factors using the New York City network of EHR.

### **Assessing Screening Efficacy in the Presence of Cancer Overdiagnosis**

♦*Ying Huang and Ziding Feng*

Fred Hutchinson Cancer Center  
yhuang@fredhutch.org

Cancer screening allows detection of cancer early in its development when treatment is most effective. Screening also carries the potential harm of overdiagnosis, namely a diagnosis made by screening that would not have caused symptoms or death during a patient's lifetime. In this project, we aim to address an important open question in screening efficacy assessment: how to choose primary endpoints and inferential procedures to efficiently account for potential overdiagnosis in screening trials, motivated by the need to design and analyze a phase IV Early Detection Initiative (EDI) trial for evaluating a pancreatic cancer screening strategy. We propose two new approaches to assess screening efficacy based on cancer stage-shift, which address potential overdiagnosis through: i) borrowing

## Abstracts

---

information about clinical diagnosis from the control arm without screening, and ii) sensitivity analysis conditional on a conservative bound on magnitude of overdiagnosis. Analytically and through extensive simulation studies, we demonstrate the advantages of the proposed approaches in estimating and testing screening efficacy compared to existing methods that either ignore overdiagnosis or adopt a valid yet conservative cumulative incidence endpoint. We also apply the proposed approaches to an ovarian cancer data application.

### **Learning Directed Acyclic Graphs for Ligands and Receptors Based on Spatially Resolved Transcriptomic Analysis**

*Pei Wang*

Icahn School of Medicine at Mount Sinai, New York, NY  
pei.wang@mssm.edu

A critical bottleneck step towards understanding the immune activation and suppression mechanisms in tumor samples is to identify transcriptional signals underlying the cell-cell communication between tumor cells and immune/stromal cells in tumor microenvironments (TME). Cell-to-cell communication extensively relies on interactions between secreted ligands and cell-surface receptors, which create a highly connected signaling network through many ligand-receptor paths. The latest advance in *in-situ* omics analyses, such as spatial transcriptomic (ST) analysis, provide unique opportunities to directly characterize ligand-receptor signaling networks that powers cell-cell communication, which has not been feasible based on either bulk or single-cell omics data. In this paper, we focus on high grade serous ovarian cancer (HGSC), and propose a novel statistical method to characterize the ligand-receptor interaction networks between adjacent tumor and stroma cells in ovarian tumors based on spatial transcriptomic data. We utilized a directed acyclic graph (DAG) model with a novel approach to handle the zero-inflated distribution observed in the ST data. It also leverages existing ligand-receptor regulation databases as prior information, and employs a bootstrap aggregation strategy to achieve robust network estimation. We applied the method to ST datasets of tumor samples from four HGSC patients, and identified common and distinct ligand-receptor regulations between adjacent

### **Learning Directed Acyclic Graphs with Mixed Variables**

♦*Jie Peng*<sup>1</sup>, *Pei Wang*<sup>2</sup> and *Shrabanti Chowdhury*<sup>2</sup>

<sup>1</sup>University of California, Davis

<sup>2</sup>Icahn School of Medicine at Mount Sinai  
jiepeng@ucdavis.edu

In this talk, we will discuss a new tool – DAGBagM– to learn directed acyclic graphs (DAGs) with both continuous and binary nodes. DAGBagM allows for either continuous or binary nodes to be parent or child nodes. It employs a bootstrap aggregating strategy to reduce false positives in edge inference. At the same time, the aggregation procedure provides a flexible framework to robustly incorporate prior information on edges. We examine DAGBagM through simulation experiments. We also apply it to proteogenomic datasets from ovarian cancer studies for identifying protein biomarkers related to treatment response.

## **Session 23CHI31: New Developments in Missing Data Problems and Related Areas**

### **Instability of Inverse Probability Weighting Methods and a Remedy for Non-Ignorable Missing Data**

*Pengfei Li*<sup>1</sup>, *Jing Qin*<sup>2</sup> and ♦*Yukun Liu*<sup>3</sup>

<sup>1</sup>University of Waterloo

<sup>2</sup>NIH

<sup>3</sup>East China Normal University  
liuyk\_ecnu@126.com

Inverse probability weighting (IPW) methods are commonly used to analyze non-ignorable missing data under the assumption of a logistic model for the missingness probability. However, solving IPW equations numerically may involve non-convergence problems when the sample size is moderate and the missingness probability is high. Moreover, those equations often have multiple roots, and identifying the best root is challenging. Therefore, IPW methods may have low efficiency or even produce biased results. We identify the pitfall in these methods pathologically: they involve the estimation of a moment-generating function, and such functions are notoriously unstable in general. As a remedy, we model the outcome distribution given the covariates of the completely observed individuals semiparametrically. After forming an induced logistic regression model for the missingness status of the outcome and covariate, we develop a maximum conditional likelihood method to estimate the underlying parameters. The proposed method circumvents the estimation of a moment-generating function and hence overcomes the instability of IPW methods. Our theoretical and simulation results show that the proposed method outperforms existing competitors greatly. Two real data examples are analyzed to illustrate the advantages of our method. We conclude that if only a parametric logistic regression is assumed but the outcome regression

### **Evaluating Dynamic Conditional Quantile Treatment Effects with Applications in Ridesharing**

♦*Ting Li*<sup>1</sup>, *Chengchun Shi*<sup>2</sup>, *Zhaohua Lu*<sup>3</sup>, *Yi Li*<sup>3</sup> and *Hongtu Zhu*<sup>4</sup>

<sup>1</sup>Shanghai University of Finance and Economics

<sup>2</sup>London School of Economics and Political Science

<sup>3</sup>Didi Chuxing

<sup>4</sup>University of North Carolina at Chapel Hill  
tingli@mail.shufe.edu.cn

Many modern tech companies, such as Google, Uber, and Didi, utilize online experiments (also known as A/B testing) to evaluate new policies against existing ones. While most studies concentrate on average treatment effects, situations with skewed and heavy-tailed outcome distributions may benefit from alternative criteria, such as quantiles. However, assessing dynamic quantile treatment effects (QTE) remains a challenge, particularly when dealing with data from ride-sourcing platforms that involve sequential decision-making across time and space. In this paper, we establish a formal framework to calculate QTE conditional on characteristics independent of the treatment. Under specific model assumptions, we demonstrate that the dynamic conditional QTE (CQTE) equals the sum of individual CQTEs across time, even though the conditional quantile of cumulative rewards may not necessarily equate to the sum of conditional quantiles of individual rewards. This crucial insight significantly streamlines the estimation and inference processes for our target causal estimand. We then introduce two varying coefficient decision process (VCDP) models and devise an innovative method to test the dynamic CQTE. Moreover, we expand our approach to accommodate data from spatiotemporal dependent experiments and examine both conditional quantile direct and indirect effects. Finally, we apply it to three dates from a ride-sharing platform.

### **Multiply Robust Estimation for Two-Part Regression Models with Missing Semicontinuous Response**

*Qiyin Zheng* and ♦*Chunlin Wang*

Xiamen University  
wangc@xmu.edu.cn

The two-part model is widely used for analyzing semicontinuous data consisting of a mixture of excess zeros and skewed positive continuous data. The conventional estimation procedures for two-part models implicitly assume that the data are completely observed. However, this assumption may be violated due to the missing data problem encountered in many applications. To address this issue in the presence of semicontinuous response data missing at random, the inverse probability weighting, imputation and doubly robust estimators are constructed. However, their consistency crucially depend on correct specification of a single model for describing the missing mechanism or/and a single (mixture) model for imputing those missing semicontinuous response. In this paper, we further develop multiply robust estimation procedures to allow for multiple models for both the missing mechanism and imputation. We establish the multiple robustness properties of the proposed estimators in the sense that they are consistent if any one of these multiple models is correctly specified. The simulation results show the multiple robustness and desirable finite-sample performance of the proposed estimators under a variety of model settings. A real psychology data application is considered for illustration.

#### Semiparametric Inference for Quantile Regression in Imbalanced Semi-Supervised Distributed System

♦Shuyi Zhang and Yong Zhou

East China Normal University  
syzhang@fem.ecnu.edu.cn

This paper considers distributed inference for imbalanced semi-supervised data under quantile regression. Since the objective function of quantile regression is non-smooth, the usual divide-and-conquer strategy and communication-efficient algorithms for smooth functions are not applicable. In this paper, we propose two algorithms, including the method based on weighted loss function and the improved Meta method, for the cases of small-sized and large-sized unlabeled datasets, respectively. The proposed algorithms can integrate the semi-supervised data scattered on different machines, so as to realize communication-efficient distributed computing for different data types and different sample sizes. Theoretical analysis on both the asymptotic properties and optimal weights are given. The resulting estimators are proved to achieve semiparametric variance lower bounds if all specifications are correct, and be more efficient than the supervised counterpart once the model is non-degenerated. A novel perturbation resampling procedure is devised for variance estimation. We explore the finite sample performance of the proposed algorithms through extensive simulation studies. Applications to the homeless data in Log Angeles are presented.

#### Session 23CHI40: Advanced Statistical Learning Methods for Complex Data

##### Distillation Decision Tree

♦Xuetao Lu and J.jack Lee  
nudttxt@163.com

Black-box machine learning models are praised for their good prediction accuracy but criticized for their lack of interpretability. Knowledge Distillation (KD) is an emerging tool that can interpret black-box models by distilling their knowledge into transparent models. Decision trees are competitive candidates for transparent models due to their well-known interpretability advantages. However, limited theoretical and empirical understanding exists for the decision tree generated from KD. In this study, we introduce the

distillation decision tree (DDT) and establish the theoretical foundations for its structural stability. We prove that the splits of DDT can achieve stability (convergence) under mild assumptions. We also develop algorithms for stabilizing the tree induction process, along with parallel computing strategies and a marginal PCA sampling algorithm to increase computational efficiency. The simulation study and real data analysis validate our theory and algorithms and demonstrate that DDT can strike a good balance between prediction accuracy and interpretability.

##### Clustering Methods for Microbiome Data

♦Peng Liu and Zhili Qiao

Iowa State University  
pliu@iastate.edu

High-throughput sequencing technologies have greatly facilitated microbiome research and have generated a large volume of microbiome data with the potential to answer key questions regarding microbiome assembly, structure, and function. Cluster analysis aims to group features that behave similarly across treatments, and such grouping helps to highlight the functional relationships among features and may provide biological insights into microbiome networks. However, clustering microbiome data are challenging due to the sparsity and high dimensionality. We propose a model-based clustering method based on Poisson hurdle models for sparse microbiome count data. We describe an expectation–maximization algorithm and a modified version using simulated annealing to conduct the cluster analysis. Moreover, we provide algorithms for initialization and choosing the number of clusters. Simulation results demonstrate that our proposed methods provide better clustering than alternative methods under various settings. We also apply the proposed method to a sorghum rhizosphere microbiome dataset, resulting in interesting biological findings.

##### Collaborative Spectral Clustering in Attributed Networks

Pengsheng Ji

Univ. of Georgia  
psji@uga.edu

We proposed a novel spectral clustering algorithm for attributed networks, where  $n$  nodes split into  $R$  non-overlapping communities and each node has a  $p$ -dimensional meta covariate from various of formats such as text, image, speech etc.. The connectivity matrix  $W_{n \times n}$  is constructed with the adjacent matrix  $A_{n \times n}$  and co-variate matrix  $X_{n \times p}$ , and  $W = (1 - \alpha)A + \alpha K(X, X')$ , where  $\alpha \in [0, 1]$  is a tuning parameter and  $K$  is a Kernel to measure the covariate similarities. We then perform the classical  $k$ -means algorithm on the element-wise ratio matrix of the first  $K$  leading eigenvector of  $W$ . Theoretical and simulation studies showed the consistent performance under both Stochastic Block Model (SBM) and Degree-Corrected Block Model (DCBM), especially in imbalanced networks where most community detection algorithms fail.

#### Session 23CHI41: Recent Development for Causal Inference and Personalized Medicine

##### Towards Trustworthy Explanation: On Causal Rationalization

Wenbo Zhang<sup>1</sup>, Tong Wu<sup>2</sup>, Yunlong Wang<sup>2</sup>, Yong Cai<sup>2</sup> and

♦Hengrui Cai<sup>1</sup>

<sup>1</sup>University of California Irvine

<sup>2</sup>IQVIA

hengrc1@uci.edu

With recent advances in natural language processing, rationalization becomes an essential self-explaining diagram to disentangle

## Abstracts

---

the black box by selecting a subset of input texts to account for the major variation in prediction. Yet, existing association-based approaches on rationalization cannot identify true rationales when two or more snippets are highly inter-correlated and thus provide a similar contribution to prediction accuracy, so-called spuriousness. To address this limitation, we novelly leverage two causal desiderata, non-spuriousness and efficiency, into rationalization from the causal inference perspective. We formally define a series of probabilities of causation based on a newly proposed structural causal model of rationalization, with its theoretical identification established as the main component of learning necessary and sufficient rationales. The superior performance of the proposed causal rationalization is demonstrated on real-world review and medical datasets with extensive experiments compared to state-of-the-art methods.

### **Multi-Threshold Change Plane Model: Estimation Theory and Applications in Subgroup Identification**

♦ Jialiang Li<sup>1</sup>, Yaguang Li, Baisuo Jin and Michael Kosorok

<sup>1</sup>National University of Singapore

stalj@nus.edu.sg

We propose a multi-threshold change plane regression model which naturally partitions the observed subjects into subgroups with different covariate effects. The underlying grouping variable is a linear function of observed covariates and thus multiple thresholds produce change planes in the covariate space. We contribute a novel 2-stage estimation approach to determine the number of subgroups, the location of thresholds and all other regression parameters. In the first stage we adopt a group selection principle to consistently identify the number of subgroups, while in the second stage change point locations and model parameter estimates are refined by a penalized induced smoothing technique. Our procedure allows sparse solutions for relatively moderate- or high-dimensional covariates. We further establish the asymptotic properties of our proposed estimators under appropriate technical conditions. We evaluate the performance of the proposed methods by simulation studies and provide illustrations using two medical data examples. Our proposal for subgroup identification may lead to an immediate application in personalized medicine.

### **Dynamic Treatment Regimes with Infinite Horizon and Outcome-Dependent Observation Process**

♦ Xin Chen and Wenbin Lu

chenxin@amss.ac.cn

Dynamic treatment regimes are sequential decision rules assigning individualized treatments to patients based on evolving treatments and covariate history. In this talk, we will focus on dynamic treatment regimes in infinite horizon and irregular observation settings where the number of decision points diverges to infinity and the occurrence times of decision points are not regular and could be informative. We develop a new framework to evaluate the dynamic treatment regimes in a relatively general setting, and construct confidence intervals for a given policy's value with reinforcement learning techniques. Since the irregular decision making times could be informative, we also develop a Cox-type renewal process model on the occurrence times of the decision points, and based on the model, we develop an adaptive estimation procedure which leads to more efficient estimates for the policy value. Simulation studies and a real data application are conducted to illustrate the proposed methods.

## **Session 23CHI56: Statistical Modeling and Inferences for Complex Data Analysis**

### **Statistical Inferences for Complex Dependence of Multimodal Imaging Data**

♦ Jinyuan Chang<sup>1</sup>, Jing He<sup>1</sup>, Jian Kang<sup>2</sup> and Mingcong Wu<sup>1</sup>

<sup>1</sup>Southwestern University of Finance and Economics

<sup>2</sup>University of Michigan

changjinyuan@swufe.edu.cn

Statistical analysis of multimodal imaging data is a challenging task, since the data involves high-dimensionality, strong spatial correlations and complex data structures. In this paper, we propose rigorous statistical testing procedures for making inferences on the complex dependence of multimodal imaging data. Motivated by the analysis of multi-task fMRI data in the Human Connectome Project (HCP) study, we particularly address three hypothesis testing problems: (a) testing independence among imaging modalities over brain regions, (b) testing independence between brain regions within imaging modalities, and (c) testing independence between brain regions across different modalities. Considering a general form for all the three tests, we develop a global testing procedure and a multiple testing procedure controlling the false discovery rate. We study theoretical properties of the proposed tests and develop a computationally efficient distributed algorithm. The proposed methods and theory are general and relevant for many statistical problems of testing independence structure among the components of high-dimensional random vectors with arbitrary dependence structures. We also illustrate our proposed methods via extensive simulations and analysis of five task fMRI contrast maps in the HCP study.

### **A Simultaneous Likelihood Test for Joint Mediation Effects of Multiple Mediators**

♦ Wei Hao and Peter Song

University of Michigan  
weihao@umich.edu

Mediation analysis is a widely used statistical tool to study the relationship between exposure and outcome through intermediate variables. When multiple mediators are present, statistical inference on the joint mediation effect can be challenging due to the composite null hypotheses with a large number of parameter configurations. To address this issue, we propose a simultaneous likelihood ratio test that utilizes a block coordinate descent algorithm to solve the constrained likelihood under the irregular null parameter space using the Lagrange multiplier approach. We establish the asymptotic null distribution and examine the performance of the proposed joint test statistic using extensive simulations and a comparison with existing tests. The simulation results demonstrate our method controls the type I error properly and generally provides better power than existing test methods.

### **Bi-Directional Clustering via Averaged Mixture of Finite Mixtures**

♦ Guanyu Hu<sup>1</sup>, Tianyu Pan<sup>2</sup> and Weineng Shen<sup>1</sup>

<sup>1</sup>University of Missouri Columbia

<sup>2</sup>University of California Irvine  
guanyu.hu@missouri.edu

Bi-directional clustering is an approach that captures the heterogeneity of a data matrix, on both rows and columns simultaneously. It has been widely applied in a variety of fields such as genomics, economics, sports, etc., to detect the clustering effect on variable-level (column) and subject-level (row) accordingly. Yet it remains under-discovered whether such bi-directional heterogeneity can be

well defined and proved to be effective using a statistical model. In this paper, we propose a density-based bi-directional clustering approach by averaging over Mixture of Finite Mixture Models (MFMs), termed as Averaged Mixture of Finite Mixtures. Our model has proven ability to capture such heterogeneity asymptotically and provide  $n^{-1/2}$  (up to a log term) contracted estimations on both density and parameters. The proposed model is manifested to be effective and tractable using simulations and helpful in mining the statistical dependency between random variables based on the applications to a Georgia county economic dataset and an NBA dataset.

#### A Simple and Robust Model for Enrollment Projection in Clinical Trials

Xiaoxi Zhang and ♦Bo Huang

Pfizer Inc.

bo.huang@pfizer.com

Enrollment projection in clinical trials is a topic gaining attention in the statistics literature in recent years. A number of methods have been proposed in this area. Some approaches are sophisticated but complicated to implement. We aim to implement a simple and robust empiric Bayes Poisson Gamma model (PGM) that is suitable for practical use. We assume a constant and site-specific underlying enrollment rate over time, which comes from a common Gamma distribution. Choice of prior parameters is data driven. We tested the proposed PGM in a simulation study as well as a number of oncology trials with various enrollment patterns, which yield satisfactory results. Compared to a flexible nonparametric model (Zhang and Long, 2010), the PGM is associated with a narrower credible interval as a result of parametric assumptions. However, the model prediction may be off when the assumptions are substantially violated.

#### Session 23CHI66: Advances in Theory and Statistical Applications of Random Fields

##### Almost-Sure Path Properties of Operator Fractional Brownian Motion

Wensheng Wang

Hangzhou Dianzi University

wswang@aliyun.com

The multivariate Gaussian random fields with matrix-based scaling laws are widely used for inference in statistics and many applied areas. In such contexts interests are often in the continuity and rates of change of spatial surfaces in any given direction. This article analyzes the almost-sure spatial surface behavior of operator fractional Brownian motion, including multivariate fractional Brownian motion. We obtain the estimations of Fernique-type inequalities and utilize them to establish the global and the local moduli of continuity for operator fractional Brownian motion in any given direction. Our results show that the rates of change of spatial surfaces are completely determined by the self-similarity exponent. Applications to the multivariate fractional Brownian motion are investigated.

##### Multivariate Spatial Modeling Based on Conditionally Negative Definite Functions

♦Juan Du<sup>1</sup> and Xiaoxi Li<sup>2</sup>

<sup>1</sup>Kansas State University

<sup>2</sup>Kansas State university  
dujuan@ksu.edu

Both covariance matrix and variogram matrix functions are used to describe the dependence structure of a multivariate random field

with second-order increments. We will explore how a conditionally negative definite matrix (function) (CNDM) can serve as an important tool in constructing multivariate covariance or variogram matrix functions. The natural connection of CNDM with both variance and covariance- based variogram matrix function is characterized. Several classes of flexible multivariate spatial and spatio-temporal covariance models are derived using the CNDM techniques and extension of Schoenberg's Theorem. Some applications of these models are illustrated using simulation studies.

##### Error Bounds for the Measurement of Random Fields and the Relation to the Statistical Model

Tianshi Lu

Wichita State University, Kansas, USA

tianshi\_lu@yahoo.com

In this talk, we present some recent results on the error estimates for the measurement of an isotropic Gaussian random field on compact two-point homogeneous spaces such as spheres. In particular, we obtained the error bounds for the interpolation of measurement based on the high-frequency behavior of the angular power spectrum of the random field. We also established some results on the smoothness of its sample paths.

##### Statistical Analysis of Multivariate Gaussian Random Fields

Yimin Xiao

xiaoy@msu.edu

In recent years, a number of classes of new multivariate random fields have been constructed by using the approaches of covariance matrices, variogram matrices, the convolution method, spectral representations, or systems of stochastic partial differential equations (SPDEs), and have been applied for modeling multivariate spatial data. In this talk, we present some recent results on the estimation and prediction of several classes of multivariate Gaussian random fields including multivariate Matern Gaussian fields, operator fractional Brownian motion, and vector-valued operator-scaling random fields. These results illustrate explicitly the effects of the dependence structures among the coordinate processes on statistical properties of multivariate Gaussian random fields.

#### Session 23CHI67: Recent Statistical Methodological Developments in Genomics

##### Zeroinflated Poisson Models with Measurement Error in the Response

♦Qihuang Zhang<sup>1</sup> and Grace Y. Yi<sup>2</sup>

<sup>1</sup>McGill University

<sup>2</sup>University of Western Ontario  
qihuang.zhang@mcgill.ca

Zero-inflated count data are common in genomics studies, where a mixture model combining a Poisson distribution with excess zeros is often used for analysis. However, measurement error in count responses presents a significant challenge to the analysis of such data. In this talk, we propose a new measurement error model for count data that includes error-contaminated count responses. We show that ignoring the measurement error effects can lead to invalid inference results, and we propose a Bayesian method to address the effects of measurement error under the zero-inflated Poisson model. We develop a data-augmentation algorithm that is easy to implement and conduct simulation studies to evaluate the method's performance. We will illustrate the proposed method's application to analyze data from a prostate adenocarcinoma genomic study.

## Abstracts

---

### Neural Network Models for Sequence-Based Tcr and Hla Association Prediction

♦*Si Liu, Philip Bradley and Wei Sun*

Fred Hutchinson Cancer Center  
sliu3@fredhutch.org

A T cell relies on its T cell receptor (TCR) to recognize foreign antigens presented by a human leukocyte antigen (HLA), which is the human version of major histocompatibility complex (MHC). HLA is the most polymorphic locus in human genome. We explore the capacity of neural networks to predict the association between HLA and TCR, based on their amino acid sequences. We quantify the functional similarities of HLA alleles based on the predictions of TCR-HLA associations, and demonstrate the association of such similarities with survival outcome of cancer patients who received immune checkpoint blockade (ICB) treatment.

### Association Analysis Between the t-Cell Receptor Repertoire and Clinical Phenotypes

*Meiling Liu<sup>1</sup>, Juna Goo<sup>2</sup>, Yang Liu<sup>3</sup>, Wei Sun<sup>1</sup>, Michael Wu<sup>1</sup>, Li Hsu<sup>1</sup> and ♦Qianchuan He<sup>1</sup>*

<sup>1</sup>Fred Hutchinson Cancer Center

<sup>2</sup>Boise State University

<sup>3</sup>Wright State University  
qhe@fredhutch.org

T cell receptors (TCRs) play critical roles in adaptive immune responses, and recent advances in genome technology have made it possible to examine the TCR repertoire at the population level. We introduce an analysis tool, TCR-L, for evaluating the association between the TCR repertoire and clinical phenotypes. The TCR-L can accommodate features that can be extracted from the TCR sequences as well as features that are hidden in the TCR sequences. Simulation studies show that the proposed approach has well controlled type I errors and good power to identify associations between TCR repertoire and disease outcomes. An application of the proposed approach to a cancer study will be shown as well.

### Knockofftrio: a Knockoff Framework for the Identification of Putative Causal Variants in Genome-Wide Association Studies with Trio Design

♦*Yi Yang<sup>1</sup>, Chen Wang<sup>2</sup>, Linxi Liu<sup>3</sup>, Joseph Buxbaum<sup>4</sup>, Zihuai He<sup>5</sup> and Iuliana Ionita-Laza<sup>2</sup>*

<sup>1</sup>City University of Hong Kong

<sup>2</sup>Columbia University

<sup>3</sup>University of Pittsburgh

<sup>4</sup>Icahn School of Medicine at Mount Sinai

<sup>5</sup>Stanford University

yi.yang@cityu.edu.hk

Family-based designs can eliminate confounding due to population substructure and can distinguish direct from indirect genetic effects, but these designs are underpowered due to limited sample sizes. Here, we propose KnockoffTrio, a statistical method to identify putative causal genetic variants for father-mother-child trio design built upon a recently developed knockoff framework in statistics. KnockoffTrio controls the false discovery rate in the presence of arbitrary correlations among tests and is less conservative and thus more powerful than the conventional methods that control the family-wise error rate via Bonferroni correction. Furthermore, KnockoffTrio is not restricted to family-based association tests and can be used in conjunction with more powerful, potentially nonlinear models to improve the power of standard family-based tests. We show, using empirical simulations, that KnockoffTrio can prioritize causal variants over associations due to linkage disequilibrium and

can provide protection against confounding due to population stratification. In applications to 14,200 trios from three study cohorts for autism spectrum disorders, we show that KnockoffTrio can identify multiple significant associations that are missed by conventional tests applied to the same data.

### Session 23CHI8: Study Design and Statistical Considerations in Oncology Development: Methodology and Case Sharing

#### Design Considerations for Early-Phase Clinical Trials of Car-T Therapies

*Wentian Guo*

AstraZeneca R&D China  
wentian.guo@astrazeneca.com

Chimeric antigen receptor (CAR) T therapy has shown great potency in cancer treatment, and the clinical development has been extended from hematopoietic malignancies to solid tumor. CAR-T therapy is a living drug, and therefore responders may benefit for a relatively longer time compared to other types of therapies. However, the complex individualized manufacturing process of CAR-T is usually associated with relative long pre-treatment and high cost. Besides, CAR-T therapy efficacy may not increase with dose, which means that traditional toxicity-only escalation methods may not be the optimal choice for CAR-T therapies. In this presentation, we will discuss the general statistical consideration of CAR-T therapies early phase clinical trials, with the focus on application of efficacy-assisted escalation methods in CAR-T trials.

#### Statistical and Operational Consideration in Umbrella Trial – a Real Case Study

♦*Monica Li, Keisuke Tada and Rick Zhang*  
Monica.li@sanofi.com

An umbrella trial is designed to evaluate multiple investigational drugs administered as single drugs or as combination therapies in a single disease population. Sub-studies can include dose-finding components to identify recommended next phase doses of an investigational drug combination before proceeding with an activity-estimating component. As a novel trial design, an umbrella trial allows efficient and accelerated drug development but also brings operational and statistical challenges. This presentation will share statistical and operational considerations in an umbrella trial based on learnings from a real case study.

#### An Oncology Case Example of Dose Optimization using the Pick-the-Winner Approach

♦*Liang Zhao<sup>1</sup>, Qiuyan Wang<sup>1</sup> and Gu Mi<sup>2</sup>*

<sup>1</sup>Department of Biostatistics and Programming, China, Sanofi, Inc.

<sup>2</sup>Department of Biostatistics and Programming, US, Sanofi, Inc.

liang2.zhao@sanofi.com

Advances of precision medicine led to many modern anticancer agents in development are targeted therapies. In contrast of traditional cytotoxic chemotherapies, targeted therapies may present a “plateau effect” between doses and efficacy where a higher dose may be associated with marginal increase in efficacy while toxicity could be substantially added, and patients may stay on these therapies for long periods of time, increasing the risk of intolerance. Therefore, new approaches to optimize doses of targeted anticancer agents are needed. Within this context, FDA has initiated Project OPTIMUS in 2021 to guide dose optimization and recommended phase 2 dose (RP2D) determination. This presentation will introduce the pick-the-winner approach within the three-outcome deci-

sion making (3ODM) framework for dose optimization, and the implementation of the approach to a Sanofi early phase oncology trial.

#### **Adjusting Survival Time Estimates Accounting for Treatment Switching**

♦*Yujuan Gao, Songzi Li and Jiang Li*

BeiGene, Ltd.

yujuan.gao@beigene.com

TBC

#### **Session 23CHI97: Dealing with Missing Data: Recent Methodological Advances**

##### **Doubly Robust Estimators for Generalizing Treatment Effects on Survival Outcomes from Randomized Controlled Trials to a Target Population**

*Dasom Lee<sup>1</sup>, ♦Shu Yang<sup>1</sup> and Xiaofei Wang<sup>2</sup>*

<sup>1</sup>North Carolina State University

<sup>2</sup>Duke University  
syang24@ncsu.edu

In the presence of heterogeneity between the randomized controlled trial (RCT) participants and the target population, evaluating the treatment effect solely based on the RCT often leads to biased quantification of the real-world treatment effect. To address the problem of lack of generalizability for the treatment effect estimated by the RCT sample, we leverage observational studies with large samples that are representative of the target population. This paper concerns evaluating treatment effects on survival outcomes for a target population and considers a broad class of estimands that are functionals of treatment-specific survival functions, including differences in survival probability and restricted mean survival times. Motivated by two intuitive but distinct approaches, i.e., imputation based on survival outcome regression and weighting based on inverse probability of sampling, censoring, and treatment assignment, we propose a semiparametric estimator through the guidance of the efficient influence function. The proposed estimator is doubly robust in the sense that it is consistent for the target population estimands if either the survival model or the weighting model is correctly specified, and is locally efficient when both are correct. In addition, as an alternative to parametric estimation, we employ the nonparametric method of sieves for flexible and robust estimation.

##### **Doubly Robust Estimation with Outliers under Missing at Random**

♦*Jae-Kwang Kim<sup>1</sup>, Jeongsup Han<sup>2</sup>, Hengfang Wang<sup>3</sup> and Youngjo Lee<sup>2</sup>*

<sup>1</sup>Iowa State University

<sup>2</sup>Seoul National University

<sup>3</sup>Fujian Normal University  
jkim@iastate.edu

Imputation is a popular technique for handling missing data. By properly incorporating the auxiliary variables, imputation can reduce the nonresponse bias and obtain efficient estimation. However, the correct specification of the statistical model may be challenging in the presence of missing data. Finding a robust estimation method that is less sensitive to failure of the assumed model is an important practical problem in the missing data literature. If the missing mechanism is missing-at-random, the doubly robust estimator is attractive since the consistency of the estimator is guaranteed either the outcome regression model or the propensity score model is correctly specified. To obtain a doubly robust estimator, we propose a unified approach of using the information projection

under the indirect model calibration condition for propensity score weighting. The resulting propensity score estimator can be equivalently expressed as an imputation estimator under the internal bias calibration condition in the modeling approach. The correct specification of the outcome regression model and the propensity score model is equivalent to that of the mean and dispersion in the double hierarchical generalized linear model. In addition, we discuss how to allow robust inference on outliers. The simulation study shows that the proposed method allows robust inference

##### **Marginal Treatment Effect Estimation without Ignorability using Observational Study**

*Guoliang Ma<sup>1</sup> and ♦Cindy Yu<sup>2</sup>*

<sup>1</sup>Iowa State University

<sup>2</sup>Iowa State University  
cindyuu@iastate.edu

Most of causal inferences (CI) assume missing at random assumption, i.e. the treatment selection probability depends on covariates only. However in real applications, it is often found that the selection probability also depends on potential outcomes, which leads to incorrect treatment effect estimation if this selection bias is ignored. Under this missing not at random (MNAR) scenario, traditional propensity score (PS) estimators fail because estimation of the propensity scores requires observed potential outcomes which are not available at the same time in the framework of CI. In this paper, we propose a new iterative estimating-solving (ES) algorithm to impute potential outcomes via semiparametric quantile regression under MNAR assumption. Model identification under MNAR is carefully discussed. We validate the proposed estimator through theories of convergence and large sample properties, and simulation studies. Variance estimation is also provided. The proposed method is applied to a real data application.

#### **Session 23CHI103: Modern Topics in Design of Experiments (DOE)**

##### **Thompson Sampling with Discrete Prior**

*Xueru Zhang<sup>1</sup>, Lan Gao<sup>2</sup> and ♦Wei Zheng<sup>2</sup>*

<sup>1</sup>Purdue University

<sup>2</sup>University of Tennessee  
wzheng9@utk.edu

Thompson sampling is a popular algorithm for multi-armed bandit problems, but its Bayesian posterior update can be computationally expensive for complex reward distributions. Recently, prior discretization has been proposed to address this issue. In this paper, we propose a new prior discretization method that guarantees the same regret rate without requiring the unreasonable assumption that the true value of the parameter is one of the discrete points. Additionally, we introduce a modified posterior update approach that further improves the performance of discrete prior Thompson sampling. We prove that the accumulated regret has  $O(\log(T))$  convergence rate with high probability. In addition, we conduct numerical experiments to validate our theoretical analysis and demonstrate that the proposed algorithm outperforms both the standard discrete prior method and the Laplace approximation approach for the continuous prior.

##### **A Maximin p-Efficient Design for Multivariate Generalized Linear Models**

♦*Yiou Li<sup>1</sup>, Lulu Kang<sup>2</sup> and Xinwei Deng<sup>3</sup>*

<sup>1</sup>Associate Prof. DePaul University

<sup>2</sup>Associate Prof. Illinois Institute of Technology

## Abstracts

---

<sup>3</sup>Prof. Virginia Tech  
yli139@depaul.edu

Experimental designs for a generalized linear model (GLM) often depend on the specification of the model, including the link function, the predictors, and unknown parameters, such as the regression coefficients. To deal with the uncertainties of these model specifications, it is important to construct optimal designs with high efficiency under such uncertainties. Existing methods such as Bayesian experimental designs often use prior distributions of model specifications to incorporate model uncertainties into the design criterion. Alternatively, one can obtain the design by optimizing the worst-case design efficiency with respect to the uncertainties of model specifications. In this work, we propose a new Maximin  $\Phi_p$ -Efficient (or Mm- $\Phi_p$  for short) design which aims at maximizing the minimum  $\Phi_p$ -efficiency under model uncertainties. Based on the theoretical properties of the proposed criterion, we develop an efficient algorithm with sound convergence properties to construct the Mm- $\Phi_p$  design. The performance of the proposed Mm- $\Phi_p$  design is assessed through several numerical examples.

### A New and Flexible Design Construction for Orthogonal Arrays for Modern Applications

*Yuanzhen He<sup>1</sup>, ♦C. Devon Lin<sup>2</sup> and Fasheng Sun<sup>3</sup>*

<sup>1</sup>School of Statistics, Beijing Normal University

<sup>2</sup>Department of Mathematics and Statistics, Queen's University

<sup>3</sup>3KLAS and School of Mathematics and Statistics, Northeast Normal University  
devon.lin@queensu.ca

Orthogonal array, a classical and effective tool for collecting data, has been flourished with its applications in modern computer experiments and engineering statistics. Driven by the wide use of computer experiments with both qualitative and quantitative factors, multiple computer experiments, multi-fidelity computer experiments, cross-validation and stochastic optimization, orthogonal arrays with certain structures have been introduced. Sliced orthogonal arrays and nested orthogonal arrays are examples of such arrays. This article introduces a flexible, fresh construction method which uses smaller arrays and a special structure. The method uncovers the hidden structure of many existing fixed-level orthogonal arrays of given run sizes, possibly with more columns. It also allows fixed-level orthogonal arrays of nearly strength three to be constructed, which are useful as there are not many construction methods for fixed-level orthogonal arrays of strength three, and also helpful for generating Latin hypercube designs with desirable low-dimensional projections. Theoretical properties of the proposed method are explored. As by-products, several theoretical results on orthogonal arrays are obtained.

### Maximum One-Factor-at-a-Time Designs for Screening in Computer Experiments

*♦Qian Xiao<sup>1</sup>, Roshan Joseph<sup>2</sup> and Douglas Ray<sup>3</sup>*

<sup>1</sup>University of Georgia

<sup>2</sup>Georgia Institute of Technology

<sup>3</sup>US Army DEVCOM Armaments Center  
xiaoqian1990v@gmail.com

Identifying important factors from a large number of potentially important factors of a highly nonlinear and computationally expensive black box model is a difficult problem. Morris screening and Sobol' design are two commonly used model-free methods for doing this. In this article, we establish a connection between these two seemingly different methods in terms of their underlying experimental design structure and further exploit this connection to develop an

improved design for screening called Maximum One-Factor-At-A-Time (MOFAT) design. We also develop efficient methods for constructing MOFAT designs with a large number of factors. Several examples are presented to demonstrate the advantages of MOFAT designs compared to Morris screening and Sobol' design methods.

### Session 23CHI104: Applications of Modern Statistical Methods for High-Dimensional and Complex Data

#### Bayesian Borrowing with Multiple Heterogeneous Historical Studies using Order Restricted Normalized Power Prior

*Zifei Han*

University of International Business and Economics  
zifeihan@uibe.edu.cn

The recent U.S. Food and Drug Administration guidance on complex innovative trial designs propagates the use of Bayesian strategies to incorporate historical information according to clinical expertise and data similarity. Meanwhile, multiple data sources from previous studies with similar settings are often eligible for historical borrowing. Though a few classes of informative priors can leverage historical information according to data compatibility in a semi-automatic manner, it is common that some exogenous factors, such as the year of the study conduct, can also influence the relevance of each historical study to the current trial. As a result, a natural ordering among the historical trials arises a priori, while currently, most informative priors fail to account for this restriction. In this talk I will introduce the order restricted normalized power prior, which guarantees a targeted order constraint on the power parameters and preserves data-adaptive borrowing. When the natural ordering is not explicit, a propensity score-based procedure is proposed to obtain a working ordering, which unifies the baseline-adaptive and the data-adaptive borrowing scheme. I will also introduce an efficient importance sampling procedure for normalization and use two clinical data sets for illustrative purposes.

#### On Sufficient Variable Screening using log Odds Ratio Filter

*Baoying Yang, ♦Wenbo Wu and Xiangrong Yin*

wenbo.wu@utsa.edu

For ultrahigh-dimensional data, variable screening is an important step to reduce the scale of the problem, hence, to improve the estimation accuracy and efficiency. In this paper, we propose a new dependence measure which is called the log odds ratio statistic to be used under the sufficient variable screening framework. The sufficient variable screening approach ensures the sufficiency of the selected input features in modeling the regression function and is an enhancement of existing marginal screening methods. In addition, we propose an ensemble variable screening approach to combine the proposed fused log odds ratio filter with the fused Kolmogorov filter to achieve supreme performance by taking advantages of both filters. We establish the sure screening properties of the fused log odds ratio filter for both marginal variable screening and sufficient variable screening. Extensive simulations and a real data analysis are provided to demonstrate the usefulness of the proposed log odds ratio filter and the sufficient variable screening procedure.

#### Bayesian Estimation of Malware Detection Metrics without Knowing Ground Truth

*Ambassador Negash<sup>1</sup>, Zifei Han<sup>2</sup>, Min Wang<sup>3</sup>, Shouhuai Xu<sup>4</sup> and*

*♦Keying Ye<sup>3</sup>*

<sup>1</sup>Footlock Inc.

<sup>2</sup>University of International Business and Economics

<sup>3</sup>University of Texas at San Antonio

<sup>4</sup>University of Colorado at Colorado Springs  
keying.ye@utsa.edu

Measurements of malware detection metrics are important in the research fields of cyber security. One of the challenging problems of developing such security metrics is due to the unknown (or noisy) ground truth. In this research, we use a Bayesian framework to study the predictive distribution of malware detection. Simulation studies are carried out using synthetic data with known ground truth and accuracies in estimations are compared. A real data is also applied with the proposed approach.

#### Exploring the Causal Relationship Between Geriatric Depression and Alzheimer's Disease

♦*Yuxia Zhang<sup>1</sup>, Yubai Yuan<sup>2</sup>, Fei Xue<sup>3</sup>, Qi Xu<sup>4</sup> and Annie Qu<sup>4</sup>*

<sup>1</sup>The University of Texas at San Antonio

<sup>2</sup>The Pennsylvania State University

<sup>3</sup>Purdue University

<sup>4</sup>University of California, Irvine  
yuxia.zhang@utsa.edu

Depression and Alzheimer's Disease (AD) are both prevalent diseases in older adults. Using the data sets from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, we explore the causal relationship between geriatric depression and AD. We estimate the average treatment effect of geriatric depression on AD while controlling for high-dimensional potential confounders, including DNA methylation. Moreover, we develop a novel causal mediation analysis approach to study mediation effects of potential mediators on the causal relationship between geriatric depression and AD. Based on the real data analysis results, we propose new prevention and treatment strategies for geriatric depression and AD through changing the selected confounders or mediators.

#### Session 23CHI105: Survival Analysis and Quantile Regression

##### From Conditional Quantile Regression to Marginal Quantile Estimation with Applications to Missing Data and Causal Inference

♦*Huijuan Ma, Jing Qin and Yong Zhou*  
hjma@fem.ecnu.edu.cn

It is well known that information on the conditional distribution of an outcome variable given covariates can be used to obtain an enhanced estimate of the marginal outcome distribution. This can be done easily by integrating out the marginal covariate distribution from the conditional outcome distribution. However, to date, no analogy has been established between marginal quantile and conditional quantile regression. This article provides a link between them. We propose two novel marginal quantile and marginal mean estimation approaches through conditional quantile regression when some of the outcomes are missing at random. The first of these approaches is free from the need to choose a propensity score. The second is double robust to model misspecification: it is consistent if either the conditional quantile regression model is correctly specified or the missing mechanism of outcome is correctly specified. Consistency and asymptotic normality of the two estimators are established, and the second double robust estimator achieves the semiparametric efficiency bound. Extensive simulation studies are performed to demonstrate the utility of the proposed approaches. An application to causal inference is introduced. For illustration, we apply the proposed methods to a job training program dataset.

##### Default Risk Prediction and Feature Extraction using a Penalized Deep Neural Network

♦*Cunjie Lin<sup>1</sup>, Nan Qiao<sup>1</sup>, Wenli Zhang<sup>1</sup>, Yang Li<sup>1</sup> and Shuangge Ma<sup>2</sup>*

<sup>1</sup>School of Statistics, Renmin University of China

<sup>2</sup>Department of Biostatistics, Yale University  
lincunjie@ruc.edu.cn

Online peer-to-peer lending platforms provide loans directly from lenders to borrowers without passing through traditional financial institutions. For lenders on these platforms to avoid loss, it is crucial that they accurately assess default risk so that they can make appropriate decisions. In this study, we develop a penalized deep learning model to predict default risk based on survival data. As opposed to simply predicting whether default will occur, we focus on predicting the probability of default over time. Moreover, by adding an additional one-to-one layer in the neural network, we achieve feature selection and estimation simultaneously by incorporating an L1-penalty into the objective function. The minibatch gradient descent algorithm makes it possible to handle massive data. An analysis of a real-world loan data and simulations demonstrate the model's competitive practical performance, which suggests favorable potential applications in peer-to-peer lending platforms.

##### Neural Network on Interval Censored Data with Application to the Prediction of Alzheimer's Disease

♦*Tao Sun<sup>1</sup> and Ying Ding<sup>2</sup>*

<sup>1</sup>Renmin University of China

<sup>2</sup>University of Pittsburgh  
sun.tao@ruc.edu.cn

Alzheimer's disease (AD) is a progressive and polygenic disorder that affects millions of individuals each year. Given that there have been few effective treatments yet for AD, it is highly desirable to develop an accurate model to predict the full disease progression profile based on an individual's genetic characteristics for early prevention and clinical management. This work uses data composed of all four phases of the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, including 1740 individuals with 8 million genetic variants. We tackle several challenges in this data, characterized by large-scale genetic data, interval-censored outcome due to intermittent assessments, and left truncation in one study phase (ADNIGO). Specifically, we first develop a semiparametric transformation model on interval-censored and left-truncated data and estimate parameters through a sieve approach. Then we propose a computationally efficient generalized score test to identify variants associated with AD progression. Next, we implement a novel neural network on interval-censored data (NN-IC) to construct a prediction model using top variants identified from the genome-wide test. Comprehensive simulation studies show that the NN-IC outperforms several existing methods in terms of prediction accuracy. Finally, we apply the NN-IC to the full ADNI data and successfully identify subgroups with differential progression risk

#### Session 23CHI106: Statistical Inference with Large Scale Data

##### Integrative Conformal p-Values for Out-of-Distribution Testing with Labeled Outliers

*Wenguang Sun*  
Zhejiang University  
wgsun@zju.edu.cn

## Abstracts

---

We present novel conformal inference methods for out-of-distribution testing that leverage side information from labeled outliers, which are commonly underutilized or even discarded by conventional conformal p-values. Blending inductive and transductive conformal inference strategies in a principled way, our methods are computationally efficient and can automatically take advantage of the most powerful model from a collection of one-class and binary classifiers. Then, we study how to control the false discovery rate in multiple testing with a conditional calibration strategy. Simulations with synthetic and real data show the proposed integrative conformal p-values outperforms existing methods.

### 去中心化数据分析中的若干统计学问题

Weidong Liu

Shanghai Jiao Tong University  
weidongl@sjtu.edu.cn

无线传感网络、多智能体决策等领域通常涉及去中心化数据分析。近十几年来，与去中心化数据分析相关的理论和算法在机器学习、数学优化、控制、工程等领域经历了蓬勃发展。然而，在统计学中，与去中心化数据分析相关的研究可以说是凤毛麟角。毫无疑问，统计学在这一领域理应扮演重要角色。因此，在这个报告中，我们将重点围绕基础统计量的计算、多重假设检验、统计优化等问题，介绍已有的相关研究，提出在去中心化框架下相应的统计学问题。

### A Robust Fusion-Extraction Procedure with Summary Statistics in the Presence of Biased Sources

Ruoyu Wang<sup>1</sup>, ♦Qihua Wang<sup>1</sup> and Wang Miao<sup>2</sup>

<sup>1</sup> Academy of Mathematics and Systems Science, Chinese Academy of Sciences

<sup>2</sup>Peking University  
qhwang@amss.ac.cn

Information from multiple data sources is increasingly available. However, some data sources may produce biased estimates due to biased sampling, data corruption, or model misspecification. This calls for robust data combination methods with biased sources. In this paper, a robust data fusion-extraction method is proposed. In contrast to existing methods, the proposed method can be applied to the important case where researchers have no knowledge of which data sources are unbiased. The proposed estimator is easy to compute and only employs summary statistics, and hence can be applied to many different fields, e.g., meta-analysis, Mendelian randomization, and distributed systems. The proposed estimator is consistent even if many data sources are biased and is asymptotically equivalent to the oracle estimator that only uses unbiased data. Asymptotic normality of the proposed estimator is also established. In contrast to the existing meta-analysis methods, the theoretical properties are guaranteed even if the number of data sources and the dimension of the parameter diverges as the sample size increases. Furthermore, the proposed method provides a consistent selection for unbiased data sources with probability approaching one. Simulation studies demonstrate the efficiency and robustness of the proposed method empirically. The proposed method is applied

### Sequential Data Integration under Dataset Shift

♦Ying Sheng<sup>1</sup>, Jing Qin<sup>2</sup> and Chiung-Yu Huang<sup>3</sup>

<sup>1</sup> Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences

<sup>2</sup>National Institute of Allergy and Infectious Diseases, National Institute of Health

<sup>3</sup>Department of Epidemiology & Biostatistics, University of California at San Francisco

shengying@amss.ac.cn

With the rapidly increasing availability of large-scale and high-velocity streaming data, efficient algorithms that can process data in batches without requiring expensive storage and computation resources have drawn considerable attention. An emerging challenge in developing efficient batch processing techniques is dataset shift, where the joint distribution of the collected data varies across batches. If not recognized and addressed properly, dataset shift often leads to erroneous statistical inferences when integrating data from different batches. In this paper, two shift-adjusted estimation procedures are developed for updated estimation of the parameter in the presence of dataset shift. Under prior probability shift, we can obtain parameter estimation and assess the degree of dataset shift simultaneously. We study the asymptotic properties of the proposed estimators and evaluate their performance in numerical studies and a real data example. This is a joint work with Jing Qin and Chiung-Yu Huang.

## Session 23CHI108: Joint Modeling of Multiple Types of Data in Health Studies

### Joint Modeling of Survival, Longitudinal and Recurrent Event Data for Individual Electronic Health Records: An Application in Community-Dwelling Elderly Surveillance

♦Hao Pan<sup>1</sup>, Xinyue Li<sup>2</sup>, Yang Zhao<sup>3</sup>, Hailiang Wang<sup>4</sup> and Kwok Leung Tsui<sup>5</sup>

<sup>1</sup> Shanghai Jiao Tong University School of Medicine, Shanghai Children's Medical Center

<sup>2</sup>City University of Hong Kong, School of Data Science

<sup>3</sup>Sun Yat-sen University School of Public Health (Shenzhen)

<sup>4</sup>The Hong Kong Polytechnic University, School of Design

<sup>5</sup>Virginia Tech, Department of Industrial and Systems Engineering  
panhao@scmc.com.cn

In community-dwelling elderly care, assessments on the functional decline are often conducted annually but not frequently enough to precede the possible occurrence of an acute disease or hospital admissions. Previous studies suggest that functional decline data are valuable information that can be fully leveraged to monitor elderly health and provide timely intervention. However, few studies have integrated functional decline data with mortality and health event data to study their complicated associations and depict the longitudinal trajectories in statistical prediction models. This paper proposes a hierarchical joint modeling framework to depict the direct and indirect effects of medical histories in Electronic Health Records (EHRs) on longitudinal observations of functional status and survival outcomes. We employed the non-homogeneous Poisson process (NHPP) model with Weibull intensity on individual EHRs to depict the time-varying counts of the medical event process. The joint model reveals the accelerated decline in functional status and the increased mortality risks based on the power-law growth of the health events. Our model was further applied to a retrospective study involving survival data, annually assessed Barthel index, and EHRs collected from 112 older adults (80 females; mean age=88 years, SD=6.97.1 years) in Hong Kong elderly care centers.

### Pathological Imaging-Assisted Cancer Gene-Environment Interaction Analysis

Kuangnan Fang<sup>1</sup>, Jingmao Li<sup>1</sup>, Qingzhao Zhang<sup>2</sup>, ♦Yaqing Xu<sup>3</sup> and Shuangge Ma<sup>4</sup>

<sup>1</sup>Department of Statistics and Data Science, School of Economics, Xiamen University, Xiamen

<sup>2</sup>The Wang Yanan Institute for Studies in Economics, Xiamen University

<sup>3</sup>School of Public Health, Shanghai Jiao Tong University School of Medicine

<sup>4</sup>Department of Biostatistics, Yale School of Public Health  
yaqing.xu@sjtu.edu.cn

Gene-environment (G-E) interactions have important implications for cancer outcomes and phenotypes beyond the main G and E effects. Compared to main-effect-only analysis, G-E interaction analysis more seriously suffers from a lack of information caused by higher dimensionality, weaker signals, and other factors. It is also uniquely challenged by the “main effects, interactions” variable selection hierarchy. Effort has been made to bring in additional information to assist cancer G-E interaction analysis. In this study, we take a strategy different from the existing literature and borrow information from pathological imaging data. Such data is a “byproduct” of biopsy, enjoys broad availability and low cost, and has been shown as informative for modeling prognosis and other cancer outcomes/phenotypes in recent studies. Building on penalization, we develop an assisted estimation and variable selection approach for G-E interaction analysis. The approach is intuitive, can be effectively realized, and has competitive performance in simulation. We further analyze The Cancer Genome Atlas (TCGA) data on lung adenocarcinoma (LUAD). The outcome of interest is overall survival, and for G variables, we analyze gene expressions. Assisted by pathological imaging data, our G-E interaction analysis leads to different findings with competitive prediction performance and stability.

#### **Human Disease Clinical Treatment Network for the Elderly: Analysis of the Medicare Inpatient Length of Stay and Readmission Data**

♦Hao Mei<sup>1</sup>, Ruofan Jia<sup>2</sup>, Guanzhong Qiao<sup>3</sup>, Zhenqiu Lin<sup>4</sup> and Shuangge Ma<sup>5</sup>

<sup>1</sup>School of Statistics, Renmin University of China

<sup>2</sup>The Wang Yanan Institute for Studies in Economics, Xiamen University

<sup>3</sup>Department of Orthopaedic, The First Hospital of Tsinghua University

<sup>4</sup>Center for Outcomes Research and Evaluation, Yale-New Haven Hospital

<sup>5</sup>Department of Biostatistics, Yale University  
hao.mei@ruc.edu.cn

Clinical treatment outcomes are the quality and cost targets that healthcare providers aim to improve. Most existing outcome analysis focuses on a single disease or all diseases combined. Motivated by the success of molecular and phenotypic human disease networks (HDNs), this study develops a clinical treatment network that describes the interconnections among diseases in terms of inpatient length of stay (LOS) and readmission. Here one node represents one disease, and two nodes are linked with an edge if their LOS and number of readmissions are conditionally dependent. This is the very first HDN that jointly analyzes multiple clinical treatment outcomes at the pan-disease level. To accommodate the unique data characteristics, we propose a modeling approach based on two-part generalized linear models and estimation based on penalized integrative analysis. Analysis is conducted on the Medicare inpatient data of 100,000 randomly selected subjects for the period of January 2010 to December 2018. The resulted network has 1008 edges for 106 nodes. We analyze key network properties including connectivity, module/hub, and temporal variation. Overall, this study can provide additional insight into diseases' properties and their in-

terconnections and assist more efficient disease management and healthcare resources allocation.

#### **Functional Adaptive Double-Sparsity Estimator for Functional Linear Regression Model with Multiple Functional Covariates**

♦Cheng Cao<sup>1</sup>, Jiguo Cao<sup>2</sup>, Hailiang Wang<sup>3</sup>, Kwok-Leung Tsui<sup>4</sup> and Xinyue Li<sup>1</sup>

<sup>1</sup>City University of Hong Kong

<sup>2</sup>Simon Fraser University

<sup>3</sup>The Hong Kong Polytechnic University

<sup>4</sup>Virginia Polytechnic Institute and State University  
chengcao3-c@my.cityu.edu.hk

Sensor devices have been increasingly used in engineering and health studies recently, and the captured multi-dimensional activity and vital sign signals can be studied in association with health outcomes to inform public health. The common approach is the scalar-on-function regression model, in which health outcomes are the scalar responses while high-dimensional sensor signals are the functional covariates, but how to effectively interpret results becomes difficult. In this study, we propose a new Functional Adaptive Double-Sparsity (FadDoS) estimator based on functional regularization of sparse group lasso and iterative adaptive weights, which can achieve global sparsity via functional variable selection and local sparsity via zero-subinterval identification within coefficient functions. We prove that the FadDoS estimator converges at a bounded rate and satisfies the oracle property under mild conditions. Application to a Kinect sensor study that utilized an advanced motion sensing device tracking human multiple joint movements and conducted among community-dwelling elderly demonstrates how FadDoS can effectively characterize the detailed association between joint movements and physical health assessments. The proposed method is not only effective in Kinect sensor analysis but also applicable to broader fields, where multi-dimensional sensor signals are collected simultaneously, to expand the use of sensor devices in health studies.

#### **Session 23CHI113: Advanced Analytic and Innovation in Healthcare**

##### **Recent Development for Ai/MI for Drug Discovery**

Haoda Fu

Eli Lilly and Company  
fu\_haoda@lilly.com

Abstract: In recent years, there has been a growing interest in the application of artificial intelligence (AI) and machine learning (ML) techniques in drug discovery. This is driven by the need for more efficient and effective ways of identifying and developing new drugs, as well as the increasing availability of large datasets and computing power. In this talk, we will review some of the recent developments in AI/ML for drug discovery, including the use of deep learning for drug design, the application of reinforcement learning for optimizing drug combinations, and the use of generative models for chemical synthesis. We will also discuss some of the challenges and limitations of these approaches, and their potential impact on the future of drug discovery. Overall, the use of AI/ML in drug discovery holds great promise for accelerating the development of new treatments and improving patient outcomes.

##### **A Digital Health Tool for Dynamic Patient Recruitment Prediction in Multicenter Clinical Trials**

Jianmin Chen<sup>1</sup>, Zhaowei Hua<sup>2</sup>, Qian Meng<sup>2</sup>, Truffaut-Chalet Luice, ♦Zhaoyang Teng<sup>2</sup> and Sammi Tang<sup>2</sup>

## Abstracts

<sup>1</sup>University of Connecticut

<sup>2</sup>Servier Pharmaceuticals

zhaoyang.teng@servier.com

Reliable prediction of patients screened and the time to reach target enrolled patients for clinical trial is important to support budget and resource planning, and timeline forecasting. Clinical trials can face challenges in the screening and enrollment process, including screening sufficient all comers for biomarker relevant trials and enrolling sufficient number of patients for some critical disease subtypes if they are rare, and so on. Bayesian Semiparametric Mixture Model was developed to model screening process and enrollment process to accommodate heterogeneity from disease subtypes or biomarker defined subtype. Model performance will be illustrated via simulations results

### Data Simulation to Forecast the Outcomes of the FAVOR III China Trial

*Yang Wang*

National Center for Cardiovascular Diseases

wangyang@mrbc-nccd.com

The aim of current study was to assess the feasibility of predicting the 1-year outcomes of the FAVOR III China trial using simulation of retrospectively assessed quantitative flow ratio (QFR) data obtained from the all-comers PANDA III trial. Among 2348 subjects from the PANDA III trial, angiography from 1391 patients was able to be analyzed with QFR. Each subject from the F3C was matched to a PANDA III patient according to the five baseline characteristics (age, sex, diabetes, multivessel disease, and existence of any vessel with diameter stenosis % >90% and thrombolysis in myocardial infarction flow <3) through a bootstrapping sampling process. Outcome predictions were based on these blinded baseline data. The primary endpoint was a composite of death, myocardial infarction, or revascularization at 1 year. After 10,000 simulations, the patients in the QFR-guided group were simulated to have a 1.9% (95% predictive intervals: -3.5% to -0.3%) absolute reduction of the occurrence of the primary study endpoint. In total, 72.7% simulated point estimates fell within the actual 95% CI of F3C (-4.7% to -1.4%) trial.

### To be Provided

*Rui (Sammi) Tang*

Sammi.tang@servier.com

To be provided

### Bayesian Learning of Covid-19 Vaccine Safety While Incorporating Adverse Events Ontology

*Bangyao Zhao, Yuan Zhong, Jian Kang and Lili Zhao*

University of Michigan

jiankang@umich.edu

While vaccines are crucial to end the COVID-19 pandemic, public confidence in vaccine safety has always been vulnerable. Many statistical methods have been applied to VAERS (Vaccine Adverse Event Reporting System) database to study the safety of COVID-19 vaccines. However, none of these methods considered the adverse event (AE) ontology. AEs are naturally related; for example, events of retching, dysphagia, and reflux are all related to an abnormal digestive system. Explicitly bringing AE relationships into the model can aid in the detection of true AE signals amid the noise while reducing false positives. We propose a Bayesian graph-assisted signal selection (BGrass) model to simultaneously estimate all AEs while incorporating the network of dependence between AEs. Under a fully Bayesian inference framework, we also propose a negative control approach to mitigate the reporting bias and an enrichment

approach to detecting AE groups of concern. For posterior computation, we construct an equivalent model representation and develop an efficient Gibbs sampler. We evaluate the performance of BGrass via extensive simulations. To study the safety of COVID-19 vaccines, we apply BGrass to analyze approximately 1,000,000 VAERS reports (01/01/2016 – 12/24/2021) involving more than 800 AEs. In particular, we found that blood clots (including deep

## Session 23CHI116: Development on Factorial Designs

### r-Optimal Designs for q-Ingredient Becker's Models for Experiments with Mixtures

♦*Chongqi Zhang<sup>1</sup> and Junpeng Li<sup>2</sup>*

<sup>1</sup>Professor

<sup>2</sup>Mr

cqzhang@gzhu.edu.cn

This talk investigates the R-optimal designs for Becker's models H2 and H3 for n=2. Using optimal design theory, three necessary results of saturated R-optimal designs are derived for both mixture models H2 and H3, including the variance matrix, the criterion function, and the equivalence of these two R-optimal designs. However, it is shown that the equivalence theorem cannot be satisfied for the saturated R-optimal designs. The excesssaturated R-optimal designs are studied by increasing the number of support points, and we prove the equivalence of the R-optimal designs for the two models for the q,q weighted simplex-centroid design. Finally, we verify the R-optimality and the effectiveness of the excess-saturated optimal designs by numerical results.

### The Construction and Properties of a Class of Clear Compromise Designs

♦*Qi Zhou and Xue Yang*

School of Statistics, Tianjin University of Finance and Economics  
daniel\_zhouqi@hotmail.com

Regular two-level factorial designs with clear effects are widely used in various experiments. To satisfy the requirement of these designs for practical applications, clear compromise plans have been proposed and studied. This paper aims at constructing a set of clear compromise plans. We first study the construction and properties of GMC (general minimum lower order confounding) designs. Then we utilize the foldover as a tool to propose a new method to construct clear compromise plans. At last, we study the properties of these designs.

### Research on Split-Plot Designs under the General Minimum Lower-Order Confounding Criterion

♦*Tao Sun and Shengli Zhao*

Qufu Normal University  
suntao\_1991@163.com

In this paper, we consider the regular s-level fractional factorial split-plot (FFSP) designs when the whole-plot (WP) factors are more important. The idea of general minimum lower-order confounding criterion is applied to such designs, and the general minimum lower-order confounding criterion of type WP (WP-GMC) is proposed. Using a finite projective geometric formulation, we derive explicit formulae connecting the key terms for the criterion with the complementary set. These results are applied to choose optimal FFSP designs under the WP-GMC criterion. Some two- and three-level WP-GMC FFSP designs are constructed.

### The Development of General Minimum Lower-Order Confounding Criterion in the Design of Experiments

*Zhiming Li*

Xinjiang University  
zmlj@xju.edu.cn

The design of experiments has played a fundamental role in many fields of scientific research. One of the critical problems is the choice of factorial designs. Up to now, quite a few optimality criteria have been proposed and investigated in depth, such as maximum resolution, minimum aberration, and clear effects. Zhang, Li, Zhao, and Ai (2008) first introduced an aliased effect number pattern (AENP) to reveal the aliasing information hidden in the defining relation and explore the relationship between these criteria. The AENP reflects the overall confounding between any effects in a design. Thus, the classification patterns of the abovementioned criteria could be expressed as some functions of the AENP. Based on the AENP, they introduced a general minimum lower-order confounding (GMC) criterion, which contains the basic information of all effects aliased with other effects at severity degree. The GMC designs are most suitable for the situation where the experimenter has prior knowledge of the importance of the factors. The AENP and GMC criterion has been developed in regular, blocked, split-plot, and mixed-level designs.

### Session 23CHI117: Stochastic Modeling of Complex Data

#### Joint Sum-Max Limit for a Class of Long-Range Dependent Processes with Heavy Tails

*Shuyang Bai and ♦ He Tang*  
University of Georgia  
ht11145@uga.edu

We consider a class of stationary processes exhibiting both long-range dependence and heavy tails. Separate limit theorems for sums and for extremes have been established recently in literature with novel objects appearing in the limits. In this article, we establish the joint sum-max limit theorems for this class of processes. In the finite-variance case, the limit consists of two independent components: a fractional Brownian motion arising from the sum, and a long-range dependent random sup measure arising from the maximum. In the infinite-variance case, we obtain in the limit two dependent components: a stable process and a random sup measure whose dependence structure is described through the local time and range of a stable subordinator. For establishing the limit theorem in the latter case, we also develop a joint convergence result for the local time and range of subordinators, which may be of independent interest.

#### How to Identify your Most Valuable Customers

*Chen Xing*  
Zeta Technologies  
chenx529@gmail.com

This presentation will focus on the fundamental concept of customer lifetime value (CLV) and its significance in quantitative marketing. The session will begin with an introduction to CLV and the various models used to calculate it, such as the Pareto/NBD model, and their practical applications. The primary goal of the presentation is to showcase how to determine the most valuable customers using CLV and the ways it can be used to make informed marketing decisions. It will also cover how CLV can aid in segmenting customer groups based on their value and designing targeted marketing campaigns accordingly. Furthermore, the presentation will feature several case studies that will highlight how CLV can be put into action in real-world business scenarios, and the value it can deliver in terms of customer retention and profitability. By the end of this session, attendees will have a comprehensive understanding of CLV, its

importance in the marketing mix, and how it can help organizations to enhance their customer engagement and retention strategies.

#### Influence Effects in Social Networks: Inward and Outward Spillovers of One Unit's Treatment

*♦ Fei Fang and Laura Forastiere*  
Yale University  
fei.fang@yale.edu

In a connected social network, users may have varying levels of influence on others when they themselves receive interventions. For example, giving an advertisement to a more influential person can have on average a greater impact on others' purchase decisions than impacts by non-celebrities. Understanding and evaluating these effects can provide valuable insights for various applications such as targeting strategies in marketing. We define the effect of one unit's treatment on the outcome of their network neighbors in two ways: i) the inward average spillover effect on a unit's outcome of a neighbor's treatment, and ii) the outward average spillover of a unit's treatment on their neighbors' outcomes. In both causal effects, we marginalize over the distribution of the treatment vector in the rest of the network under a hypothetical Bernoulli trial. We investigate the comparison between the two causal effects in directed networks with different properties, including the conditions under which they are equivalent. We present examples which satisfy these conditions. Additionally, we develop Horvitz – Thompson estimators for both types of causal effects and their design-based variance in randomized experiments.

#### High-Quantile Regression for Tail-Dependent Time Series

*Ting Zhang*  
University of Georgia  
tingzhang@uga.edu

Quantile regression is a popular and powerful method for studying the effect of regressors on quantiles of a response distribution. However, existing results on quantile regression were mainly developed for cases in which the quantile level is fixed, and the data are often assumed to be independent. Motivated by recent applications, we consider the situation where (i) the quantile level is not fixed and can grow with the sample size to capture the tail phenomena, and (ii) the data are no longer independent, but collected as a time series that can exhibit serial dependence in both tail and non-tail regions. To study the asymptotic theory for high-quantile regression estimators in the time series setting, we introduce a tail adversarial stability condition, which had not previously been described, and show that it leads to an interpretable and convenient framework for obtaining limit theorems for time series that exhibit serial dependence in the tail region, but are not necessarily strongly mixing. Numerical experiments are conducted to illustrate the effect of tail dependence on high-quantile regression estimators, for which simply ignoring the tail dependence may yield misleading p-values.

### Session 23CHI119: Advanced Statistical Methods for Data Analysis

#### Combining Extreme Value Theory with Martingale Regression in Market Risk Analytics and Portfolio Management.

*♦ Wei Dai and Tze Leung Lai*  
daiwei1969@gmail.com

This article introduces a new econometric approach to the estimation of VaR and convex risk measures in financial risk management, which uses a martingale regression for asset pricing and the extreme value theory (EVT) for the martingale difference residuals.

## Abstracts

We show that the MLGEV (maximum likelihood for generalized extreme value distributions) and the peaks over threshold (POT) method in EVT can apply to the dynamically scaled residuals. This approach, therefore, addresses the pitfalls of EVT techniques while enabling them to realize their widely recognized potential for estimating extreme quantiles and probabilities.

### A CLT for the Lss of Large Dimensional Sample Covariance Matrices with Diverging Spikes

♦Zhijun Liu, Jiang Hu, Zhidong Bai and Haiyan Song

Northeast Normal University  
2049767761@qq.com

In this paper, we establish the central limit theorem (CLT) for linear spectral statistics (LSS) of large-dimensional sample covariance matrix when the population covariance matrices are not uniformly bounded, which is a nontrivial extension of the Bai-Silverstein theorem (BST) (2004). The latter has strongly influenced the development of high-dimensional statistics, especially in applications of random matrix theory to statistics. However, the assumption of uniform boundedness of the population covariance matrices in BST is not satisfied in some fields, such as in economics, the variances of principal components can grow to infinity (see pervasiveness assumption B in Bai and Ng 2002). The aim of this paper is to remove the barriers for the applications of the BST. The new CLT allows spiked eigenvalues to exist, which may be bounded or tend to infinity. An important feature of our result is that the variance in CLT is related to both spiked eigenvalues and the bulk eigenvalues, and which part dominates depending on which variance is nonnegligible in the summation of the variances.

### Inference in Nonstationary Heavy-Tailed AR Process via Model Selection

♦Feifei Guo<sup>1</sup> and Rui She<sup>2</sup>

<sup>1</sup>Beijing Institute of Technology

<sup>2</sup>Southwestern University of Finance and Economics  
feifeiguo@bit.edu.cn

This paper develops a Lasso-based procedure to do statistical inference in the autoregressive (AR) process with heavy-tailed heteroscedastic noises. We first study the traditional adaptive Lasso estimator to the nonstationary coefficient, and show that it can distinguish between stationary and nonstationary autoregressions by detecting accurately whether the coefficient is zero with probability approaching to one. After the unit-root testing, we consider the self-weighted least absolute deviation estimator (SWLAD) with the adaptive Lasso penalty to the stationary coefficients. As expected, the penalized SWLAD achieves the oracle properties. This means that it is consistent, selects the correct sparsity pattern and estimates the coefficients belonging to the relevant variables at the same asymptotic efficiency as if only these had been included in the model. The whole procedure does not rely on any priori information on the lag order, the stationary coefficients and the tail index. This makes it appealing in practice, with wide applications in finance and econometrics. A simulation study is carried out to assess the performance of our method, and two real examples are provided to demonstrate its applicability.

### A Weighted Average Distributed Estimator for High Dimensional Parameter

♦Jun Lu<sup>1</sup>, Mengyao Li<sup>2</sup> and Chenping Hou<sup>1</sup>

<sup>1</sup>National University of Defense Technology

<sup>2</sup>Xi'an Jiaotong University  
penguin1020@foxmail.com  
TBC

### Session 23CHI122: Addressing Computational Challenges in Analyzing High-Throughput Genomic and Epigenetics Data

#### G3dc: a Gene-Graph-Guided Selective Deep Clustering Method for Single Cell Rna-Seq Data

Shuqing He, Jicong Fan and ♦Tianwei Yu  
yutianwei@cuhk.edu.cn

Single-Cell RNA sequencing (scRNA-seq) technology measures the expression of thousands of genes at the cellular level. Analyzing single cell transcriptome allows the identification of heterogeneous cell groups, cellular-level regulations, and the trajectory of cell development. An important aspect in the analyses of scRNA-seq data is the clustering of cells, which is hampered by issues such as high dimensionality, cell type imbalance, redundancy, and dropout. Given cells of each type are functionally consistent, incorporating biological relations between genes may improve the clustering results. Here, we develop a deep embedded clustering method, G3DC, that incorporates a graph regularization based on the existing gene network and a feature selector based on the  $\ell_{2,1}$ -norm regularization, together with a reconstruction loss to achieve both discriminative and informative embedding. The involvement of the gene network strengthens clustering performance, while helping the selection of functionally coherent genes that contribute to the clustering results. Extensive experiments have shown that G3DC offers high clustering accuracy with regard to agreement with true cell types, outperforming other leading single-cell clustering methods. In addition, G3DC selects biologically relevant genes that contribute to the clustering, providing insight into biological functionality that differentiate cell groups.

#### Supervised Cell Type Identification for Single Cell Atac-Seq Data

Hao Wu

Shenzhen Institute of Advanced Technology  
wuhao@siat.ac.cn

Computational cell type identification (celltyping) is a fundamental step in single-cell omics data analysis. Supervised celltyping methods have gained increasing popularity in single-cell RNA-seq data because of the superior performance and the availability of high-quality reference datasets. Recent technological advances in profiling chromatin accessibility at single-cell resolution (scATAC-seq) have brought new insights to the understanding of epigenetic heterogeneity and gene regulatory mechanism. With continuous accumulation of scATAC-seq datasets, supervised celltyping method specifically designed for scATAC-seq is in urgent need. In this talk, I will present our recent method developments on supervised celltyping for scATAC-seq using scATAC-seq data as reference. We developed Cellcano, a novel computational method based on a two-round supervised learning algorithm. The method alleviates the distributional shift between reference and target data and significantly improves the prediction performance.

#### Evaluation of Epitranscriptome-Wide N6-Methyladenosine Differential Analysis Methods

Daoyu Duan<sup>1</sup>, Wen Tang<sup>1</sup>, Runshu Wang<sup>2</sup>, ♦Zhenxing Guo<sup>3</sup> and Hao Feng<sup>1</sup>

<sup>1</sup>Department of Population and Quantitative Health Sciences, Case Western Reserve University

<sup>2</sup>Department of Biostatistics, University of Michigan

<sup>3</sup>School of Data Science, The Chinese University of Hong Kong - Shenzhen

guozhenxing@cuhk.edu.cn

RNA methylation has emerged recently as an active research domain to study post-transcriptional alteration in gene expression regulation. Various types of RNA methylation, especially N6-methyladenosine (m6A), are involved in human disease development. One of the fundamental questions in RNA methylation data analysis is to identify the Differentially Methylated Regions (DMRs), by contrasting cases and controls. Multiple statistical approaches have been recently developed for DMR detection, but there is a lack of a comprehensive evaluation for these analytical methods. Here, we thoroughly assess all eight existing methods for m6A DMR calling, using both synthetic and real data. For all methods, low sensitivities are observed among regions with low input levels, but they can be drastically boosted by an increase in sample size. TRESS and exomePeak2 perform the best using metrics of detection precision, FDR, type I error control, and runtime, though hampered by low sensitivity. Analyses on three real datasets suggest differential preference on identified DMR length and uniquely discovered regions, between these methods.

#### Rationally Design Generative-Adversarial Models for Delineating the Regulatory Map in Silico

Ge Gao

Peking University

gaog@mail.cbi.pku.edu.cn

Human individual cells, as the basic biological units of our bodies, carry out their functions through rigorous regulation of gene expression and exhibit heterogeneity among each other in every human tissue. In addition to identify individual genes, one is often interested in how multiple genes interact to form regulatory circuits and carry out cellular functions. Combining massive omics data and leading-edge statistical modeling/machine learning approaches, we have developed set of novel bioinformatic technologies to delineate the regulatory map and characterize the functional genome in action globally during past years. Here we will present our recent advances as well as their potential applications in clinical and translational study.

#### Session 23CHI171: Advances in Statistical Methods for Biomedical Studies

##### On the Instrumental Variable Estimation with many Weak and Invalid Instruments

♦Yiqi Lin<sup>1</sup>, Frank Windmeijer<sup>2</sup>, Xinyuan Song<sup>1</sup> and Qingliang Fan<sup>3</sup>

<sup>1</sup>Dept of Stat, The Chinese University of Hong Kong

<sup>2</sup>Dept of Stat, University of Oxford

<sup>3</sup>Dept of Econ, The Chinese University of Hong Kong  
YiqiLIN@link.cuhk.edu.hk

We discuss the fundamental issue of identification in linear instrumental variable (IV) models with unknown IV validity. We revisit the popular majority and plurality rules and show that no identification condition can be “if and only if” in general. With the assumption of the “sparsest rule”, which is equivalent to the plurality rule but becomes operational in computation algorithms, we investigate and prove the advantages of non-convex penalized approaches over other IV estimators based on two-step selections, in terms of selection consistency and accommodation for individually weak IVs. Furthermore, we propose a surrogate sparsest penalty that aligns with the identification condition and provides oracle sparse structure simultaneously. Desirable theoretical properties are derived for the proposed estimator with weaker IV strength conditions compared

to the previous literature. Finite sample properties are demonstrated using simulations and the selection and estimation method is applied to an empirical study concerning the effect of trade on economic growth.

##### A Flexible Bayesian Clustering of Dynamic Subpopulations in Neural Spiking Activity

Ganchao Wei, Ian Stevenson and ♦Xiaojing Wang

University of Connecticut

xiaojing.wang@uconn.edu

With advances in neural recording techniques, neuroscientists are now able to record the spiking activity of many hundreds of neurons simultaneously, and new statistical methods are needed to understand the structure of this large-scale neural population activity. Although previous work has tried to summarize neural activity within and between known populations by extracting low-dimensional latent factors, in many cases what determines a unique population may be unclear. Neurons differ in their anatomical location, but also, in their cell types and response properties. To identify populations directly related to neural activity, we develop a clustering method based on a mixture of dynamic Poisson factor analyzers (mixDPFA) model, with the number of clusters and dimension of latent factors for each cluster treated as unknown parameters. To analyze the proposed mixDPFA model, we propose a Markov chain Monte Carlo (MCMC) algorithm to efficiently sample its posterior distribution. Validating our proposed MCMC algorithm through simulations, we find that it can accurately recover the unknown parameters and the true clustering in the model, and is insensitive to the initial cluster assignments. We then apply the proposed mixDPFA model to multi-region experimental recordings, where we find that the proposed method can identify novel, reliable clusters of

##### A Knockoff Framework for Effective Biomarker Identification in Drug Development: With Application to a Psoriatic Arthritis Clinical Trial

Matthías Kormákksson, Kostas Sechidis, Xuan Zhu, David Ohlssen and ♦Cong Zhang

Novartis

cong.zhang@novartis.com

Machine learning and big data have been prompting the quantitative decision making in drug development, providing new opportunities for facilitating the development process and precision medicine. With the ability to sift through vast amounts of complex data, machine learning algorithms have made significant breakthroughs in identifying biomarkers for efficacy, prognosis, and personalized therapy. However, the challenge of finding true biomarkers that can provide actionable insights for drug development remains a major obstacle. To address this challenge, we have proposed a knockoff framework for controlling false discovery when identifying prognostic and predictive biomarkers via machine learning algorithms. We demonstrate the utility of the framework on a large clinical data pool of more than 2000 patients with psoriatic arthritis evaluated in four clinical trials, where we determine both prognostic and predictive factors of a well-established clinical outcome. Our work provides a powerful and flexible knockoff framework for variable selection and it has a wide applications to commonly encountered datasets in medical practice and other fields. With the increasing availability of large and complex datasets, knockoff-based biomarker identification is expected to play an increasingly important role in precision medicine and drug development.

##### On the Instrumental Variable Estimation with many Weak and

## Abstracts

---

### Invalid Instruments

*Yiqi LIN*  
YiqiLIN@link.cuhk.edu.hk

We discuss the fundamental issue of identification in linear instrumental variable (IV) models with unknown IV validity. We revisit the popular majority and plurality rules and show that no identification condition can be “if and only if” in general. With the assumption of the “sparsest rule”, which is equivalent to the plurality rule but becomes operational in computation algorithms, we investigate and prove the advantages of non-convex penalized approaches over other IV estimators based on two-step selections, in terms of selection consistency and accommodation for individually weak IVs. Furthermore, we propose a surrogate sparsest penalty that aligns with the identification condition and provides oracle sparse structure simultaneously. Desirable theoretical properties are derived for the proposed estimator with weaker IV strength conditions compared to the previous literature. Finite sample properties are demonstrated using simulations and the selection and estimation method is applied to an empirical study concerning the effect of trade on economic growth.

### Session 23CHI32: Advances Developments in Statistical Methodology

#### Efficient Estimation of Cox Model with Random Change Point

*Yalu Ping<sup>1</sup>, ♦Xuerong Chen<sup>1</sup> and Jianguo Sun<sup>2</sup>*

<sup>1</sup>Southwestern University of Finance and Economics

<sup>2</sup>University of Missouri  
chenxr522@foxmail.com

In clinical studies, the risk of a disease may dramatically change when some biological indexes of the human body exceed some thresholds. Furthermore, the differences in individual characteristics of patients such as physical and psychological experience may lead to subject-specific thresholds or change points. Although a large literature has been established for regression analysis of failure time data with change points, most of the existing methods assume the same, fixed change point for all study subjects. In this paper, we consider the situation where there exists a subject-specific change point and a Cox type model is presented. The proposed models also offer a framework for subgroup analysis. For inference, a sieve maximum likelihood estimation procedure is proposed and the asymptotic properties of the resulting estimators are established. An extensive simulation study is conducted to assess the empirical performance of the proposed method and indicates that it works well in practical situations. Finally the proposed approach is applied to a set of primary biliary cirrhosis data.

#### Nearest-Neighbor Sampling Based Conditional Independence Testing

*Shuai Li<sup>1</sup>, ♦Ziqi Chen<sup>1</sup>, Hongtu Zhu<sup>2</sup>, Dan Wang<sup>3</sup> and Wang Wen<sup>4</sup>*

<sup>1</sup>East China Normal University

<sup>2</sup>The University of North Carolina at Chapel Hill

<sup>3</sup>New York University Shanghai

<sup>4</sup>Central South University  
zqchen@fem.ecnu.edu.cn

The conditional randomization test (CRT) was recently proposed to test whether two random variables  $X$  and  $Y$  are conditionally independent given random variables  $Z$ . The CRT assumes that the conditional distribution of  $X$  given  $Z$  is known under the null hypothesis and then it is compared to the distribution of the observed samples of the original data. The aim of this paper is to develop a novel alternative of CRT by using nearest-neighbor sampling without assuming

the exact form of the distribution of  $X$  given  $Z$ . Specifically, we utilize the computationally efficient 1-nearest-neighbor to approximate the conditional distribution that encodes the null hypothesis. Then, theoretically, we show that the distribution of the generated samples is very close to the true conditional distribution in terms of total variation distance. Furthermore, we take the classifier-based conditional mutual information estimator as our test statistic. The test statistic as an empirical fundamental information theoretic quantity is able to well capture the conditional-dependence feature. We show that our proposed test is computationally very fast, while controlling type I and II errors quite well. Finally, we demonstrate the efficiency of our proposed test in both synthetic and real data analyses.

#### Mean Change Point Detection Based on Jump Information Criterion

*Zhiming Xia*

Northwest University  
statxzm@nwu.edu.cn

In this talk, we will introduce a new methodology, JIC(Jump Information Criterion), for change-point detection. The change-point model has been proved to have practical significance in many fields, in which the mean single change-point model is the basis of the change-point model. This talk mainly focuses on solving the statistical inference problem of whether there is a change point in a single change-point statistical model. For the mean single change-point model, this paper converts the traditional hypothesis testing idea into an estimation problem for the number of variable points to be 0 or 1, thus avoiding the critical value selection problem; the optimization objective function based on the jump information criterion is established, and the consistency of the number of change points and its convergence speed are proved, and finally the construction form of the optimal jump information criterion is derived. The numerical experimental results show that our proposed method has superior statistical performance compared to existing test-based methods.

### Session 23CHI33: Recent Statistical Developments for Complex High-Dimensional Data

#### Identification of Prognostic and Predictive Subgroups for Clustered Survival Data

*Ye He<sup>1</sup>, Dongsheng Tu<sup>2</sup> and ♦Ling Zhou<sup>3</sup>*

<sup>1</sup>Sichuan Normal University

<sup>2</sup>Queen's University

<sup>3</sup>Southwestern University of Finance and Economics  
zhouling@swufe.edu.cn

Identification of the subgroups of patients with heterogeneity in clinical outcomes or treatment effects based on baseline characteristics has been an active area of study in clinical research, especially in the field of personalized medicine. Existing methods can identify either prognostic or predictive subgroups separately based on independent clinical outcomes. In this paper, we propose a novel double-center-augmented frailty (DCAF) model to identify prognostic and predictive subgroups simultaneously based on clustered survival data. In particular, the proposed DCAF model enables simultaneous subgroup analysis and variable selection to identify prognostic subgroups, and homogeneity fusion on high-dimensional biomarkers to identify predictive subgroups. Instead of commonly used pairwise penalties, a novel differentiable center-augmented harmonic-type penalty is adopted to identify subgroups. Based on the gradient descent and Majorize-Minimize (MM) algo-

rithm, a simple to implement procedure, KHM-GA, is proposed. The double-group strategy in our approach leads to superior performance in the estimation and clustering accuracy when there exists heterogeneity across clusters and weak signals across high dimensional biomarkers. The large sample properties of the estimators from the proposed method and results from extensive simulation studies are also provided.

### A Variational Bayesian Approach to Identifying Whole-Brain Directed Networks with Fmri Data

*Yaotian Wang<sup>1</sup>, Guofen Yan<sup>2</sup>, Xiaofeng Wang<sup>3</sup>, Shuoran Li<sup>1</sup>, Lingyi Peng<sup>1</sup>, Dana Tudorascu<sup>1</sup> and ♦Tingting Zhang<sup>1</sup>*

<sup>1</sup>University of Pittsburgh

<sup>2</sup>University of Virginia

<sup>3</sup>Cleveland Clinic  
tiz67@pitt.edu

The brain is a high-dimensional directed network system as it consists of many regions as network nodes that exert influence on each other. The directed influence exerted by one region on another is referred to as directed connectivity. We aim to reveal whole-brain directed networks based on resting-state functional magnetic resonance imaging (fMRI) data of many subjects. However, it is both statistically and computationally challenging to produce scientifically meaningful estimates of whole-brain directed networks. To address the statistical modeling challenge, we assume modular brain networks, which reflect functional specialization and functional integration of the brain. We address the computational challenge by developing a variational Bayesian method to estimate the new model. We apply our method to resting-state fMRI data of many subjects and identify modules and directed connections in whole-brain directed networks. The identified modules are accordant with functional brain systems specialized for different functions. We also detect directed connections between functionally specialized modules, which is not attainable by existing network methods based on functional connectivity. In summary, this paper presents a new computationally efficient and flexible method for directed network studies of the brain as well as new scientific findings regarding the functional organization of the human brain.

### d-Gcca: Decomposition-Based Generalized Canonical Correlation Analysis for Multi-View High-Dimensional Data

♦Hai Shu<sup>1</sup>, Zhe Qu<sup>2</sup> and Hongtu Zhu<sup>3</sup>

<sup>1</sup>New York University

<sup>2</sup>Tulane University

<sup>3</sup>The University of North Carolina at Chapel Hill  
hs120@nyu.edu

Modern biomedical studies often collect multi-view data, that is, multiple types of data measured on the same set of objects. A popular model in high-dimensional multi-view data analysis is to decompose each view's data matrix into a low-rank common-source matrix generated by latent factors common across all data views, a low-rank distinctive-source matrix corresponding to each view, and an additive noise matrix. We propose a novel decomposition method for this model, called decomposition-based generalized canonical correlation analysis (D-GCCA). The D-GCCA rigorously defines the decomposition on the L2 space of random variables in contrast to the Euclidean dot product space used by most existing methods. Moreover, to well calibrate common latent factors, we impose a desirable orthogonality constraint on distinctive latent factors. Existing methods, however, inadequately consider such orthogonality and may thus suffer from substantial loss of undetected common-source variation. Our D-GCCA takes one step further than gen-

eralized canonical correlation analysis by separating common and distinctive components among canonical variables, while enjoying an appealing interpretation from the perspective of principal component analysis. Consistent estimators of our D-GCCA method are established with good finite-sample numerical performance. The superiority of D-GCCA over state-of-the-art methods is also corroborated in simulations and real-world data examples.

### Session 23CHI36: Statistical Machine Learning and Complex Life Time Data Analysis

#### Identification of Survival Relevant Genes with Measurement Error in Gene Expression Incorporated

♦Juan Xiong<sup>1</sup> and Wenqing He<sup>2</sup>

<sup>1</sup>Shenzhen University

<sup>2</sup>Western University  
jxiong@szu.edu.cn

Modern gene expression technologies, such as microarray and the next generation RNA sequencing, enable simultaneous measurement of expressions of a large number of genes, and therefore represent important tools in the personalized medicine research for improving the patient survival prediction accuracy. However, survival analysis with gene expression data can be challenging due to the high dimensionality. Proper identification of survival relevant genes is thus imperative for building suitable prediction models. In spite of the fact that gene expressions are typically subject to measurement errors introduced from the complex experimental procedure, the issue of measurement error is often ignored in survival gene identifications. In this article, the effect of measurement error on the identification of survival relevant genes is explored under the accelerated failure time model setting. Survival relevant genes are identified by regularizing the weighted least square estimator with the adaptive LASSO penalty. The simulation-extrapolation method is incorporated to adjust for the impact of measurement error on gene identification. The performance of the proposed method is assessed by simulation studies and the utility of the proposed method is illustrated by a real data set collected from the diffuse large-B-cell lymphoma study. The results show that the proposed method yields better

#### A Novel Approach to Ordinal Classification with Deep Neural Networks

♦Yiwei Fan<sup>1</sup>, Xiaoshi Lu<sup>2</sup> and Xiaoling Lu<sup>2</sup>

<sup>1</sup>School of Mathematics and Statistics, Beijing Institute of Technology

<sup>2</sup>School of Statistics, Renmin University of China  
fanyiwei@bit.edu.cn

Deep learning has enjoyed tremendous success in many fields but its application to ordinal classification remains scarce. While it is flexible to model complex data, a major concern is its lack of interpretability. In this paper, we introduce deep learning to partially linear concordance-based ordinal classification, where covariates of primary interest for interpretation are modelled linearly and all other covariates are modelled nonparametrically using a deep neural network. Statistical properties including consistency and convergence rate of both the linear coefficient estimators and the nonparametric function estimator are established. Simulations and extensive real applications on benchmark datasets demonstrate the significant advantage of the proposed method in classification accuracy.

#### Supervised Topic Modeling: Optimal Estimation and Statistical

## Abstracts

---

### Inference

♦Ruijia Wu<sup>1</sup>, Linjun Zhang<sup>2</sup> and T. Tony Cai<sup>3</sup>

<sup>1</sup>Shanghai Jiao Tong University

<sup>2</sup>Rutgers University

<sup>3</sup>University of Pennsylvania  
rjwu@sjtu.edu.cn

The rapid growth of digital textual data has made it increasingly important to develop statistical methods for the analysis of such data with theoretical guarantees. In this talk, we focus on supervised topic modeling within the framework of generalized linear models (GLMs) and probabilistic latent semantic indexing (pLSI) models. One of the major challenges of the analysis is that the covariates are unobservable. We propose a novel bias-adjusted estimator of the covariates and use it to estimate the regression vector. We establish minimax optimal rates of convergence and show that the proposed estimator is rate-optimal up to a logarithmic factor. In addition, we consider statistical inference for individual regression coefficients and construct confidence intervals based on an asymptotically unbiased and normally distributed estimator. The effectiveness of our proposed algorithms is demonstrated through simulation studies and applications to the analysis of a movie review dataset.

### Multistate Modeling and Structure Selection for Multitype Recurrent Events and Terminal Event Data

Chuoxin Ma

Beijing Normal University-Hong Kong Baptist University United International College (UIC)  
chuoxinma@uic.edu.cn

In cardiovascular disease studies, a large number of risk factors are measured but it often remains unknown whether all of them are relevant variables and whether the impact of these variables is changing with time or remains constant. In addition, more than one kind of cardiovascular disease events can be observed in the same patient and events of different types are possibly correlated. It is expected that different kinds of events are associated with different covariates and the forms of covariate effects also vary between event types. To tackle these problems, we proposed a multistate modeling framework for the joint analysis of multitype recurrent events and terminal event. Model structure selection is performed to identify covariates with time-varying coefficients, time-independent coefficients, and null effects. This helps in understanding the disease process as it can detect relevant covariates and identify the temporal dynamics of the covariate effects. It also provides a more parsimonious model to achieve better risk prediction.

### Session 23CHI37: Statistical Challenges for Analyzing Complex Data

#### Semiparametric Efficient Estimation of Genetic Relatedness with Double Machine Learning

Xu Guo<sup>1</sup>, Yiyuan Qian<sup>1</sup>, Hongwei Shi<sup>1</sup>, Weichao Yang<sup>1</sup> and ♦Niwen Zhou<sup>2</sup>

<sup>1</sup>School of Statistics, Beijing Normal University

<sup>2</sup>Advanced Institute of Natural Sciences, Beijing Normal University  
nwzhou@mail.bnu.edu.cn

In this paper, we propose semiparametric efficient estimators of genetic relatedness between two traits in a model-free framework. Most existing methods require specifying certain parametric models involving the traits and genetic variants. However, the bias due to model misspecification may yield misleading statistical results.

Moreover, the semiparametric efficient bounds for estimators of genetic relatedness are still lacking. In this paper, we develop semiparametric efficient estimators with machine learning methods and construct valid confidence intervals for two important measures of genetic relatedness: genetic covariance and genetic correlation, allowing both continuous and discrete responses. Based on the derived efficient influence functions of genetic relatedness, we propose a consistent estimator of the genetic covariance as long as one of genetic values is consistently estimated. The data of two traits may be collected from the same group or different groups of individuals. Various numerical studies are performed to illustrate our introduced procedures. We also apply proposed procedures to analyze Carworth Farms White mice genome-wide association study data.

#### Causal Inference Methods for Multiple Treatment Group Evaluations

Hongwei Zhao

Texas A&M University  
hongweizhao@tamu.edu

Causal inference methods are discussed for comparing treatment effects when multiple treatment groups are present. Key ideas of causal inference and the potential outcome framework will be reviewed, and several propensity score based methods will be considered. Additionally, machine learning methods will be applied to increase the flexibility of the model. Simulation studies will be conducted to compare the consistency and efficiency of different causal inference methods. Finally, these methods will be demonstrated using a real example where multiple treatment groups are involved.

#### Unifying Estimation and Inference for Linear Regression with Stationary and Integrated or Near-Integrated Variables

Shaoxin Hong<sup>1</sup>, Daniel Henderson<sup>2</sup>, ♦Jiancheng Jiang<sup>3</sup> and Qingshan Ni<sup>4</sup>

<sup>1</sup>Shandong University

<sup>2</sup>University of Alabama

<sup>3</sup>University of North Carolina at Charlotte

<sup>4</sup>Hunan University

jjiang1@uncc.edu

In linear time series regression, there is a discrepancy in the limiting distributions of least squares estimators for stationary and integrated or near-integrated variables. This makes statistical inference difficult in practice as it must be decided which distribution should be used prior to constructing interval estimates and conducting hypothesis tests. This motivates us to develop a multiple linear regression model with stationary and integrated or near-integrated state variables to reduce this difficulty and propose a unifying inference procedure for the coefficient estimates. To facilitate this unifying inference, we propose a weighted estimation technique. The asymptotic distributions of the proposed estimators are developed. However, the asymptotic variance cannot be estimated well in practice due to erratic behavior of the residual-based variance estimate. As traditional bootstrap interval estimates do not work either (as bootstrap sampling from the residuals in this nonstationary setting can be riddled with many large values), a random weighting bootstrap method is proposed for constructing confidence regions. The proposed method works well (with time constant or time varying error variance) in our simulations and outperforms existing approaches. In an empirical application which examines the predictability of asset returns, we further show how our methods can be implemented when some

#### Bolt-Ssi: a Statistical Approach to Screening Interaction Effects

**for Ultra-High Dimensional Data**

♦*Min Zhou<sup>1</sup>, Mingwei Dai<sup>2</sup>, Yuan Yao<sup>3</sup>, Jin Liu<sup>4</sup>, Can Yang<sup>5</sup> and Heng Peng<sup>6</sup>*

<sup>1</sup>Beijing Normal University–Hong Kong Baptist University United International College

<sup>2</sup>Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics

<sup>3</sup>Victoria University of Wellington, School of Mathematics and Statistics

<sup>4</sup>Duke-NUS Graduate Medical School

<sup>5</sup>The Hong Kong University of Science and Technology

<sup>6</sup>Hong Kong Baptist University

minzhou@uic.edu.cn

Detecting interaction effects among predictors on the response variable is a crucial step in various applications. In this paper, we first propose a simple method for sure screening interactions (SSI), SSI works well for problems of moderate dimensionality without the heredity assumption. To ultra-high dimensional problems, motivated by discretization associated Boolean representation and operations and the contingency table for discrete variables, we propose a fast algorithm, named “BOLT-SSI”. The statistical theory has been established for SSI and BOLT-SSI, guaranteeing their sure screening property. The performance of SSI and BOLT-SSI are evaluated by comprehensive simulation and real case studies. Numerical results demonstrate that SSI and BOLT-SSI can often outperform their competitors in terms of computational efficiency and statistical accuracy. The proposed method can be applied for fully detecting interactions with more than 300,000 predictors. Based on this study, we believe that there is a great need to rethink the relationship between statistical accuracy and computational efficiency. We have shown that the computational performance of a statistical method can often be greatly improved by exploring the advantages of computational

pared with the lower bound, the upper bound on the estimation error is nearly

**Reseach on Consumer Purchase Intention Based on Short Video Platform**

♦*Bo Li<sup>1</sup>, Xinghui Xiao<sup>2</sup> and Chen Wang<sup>1</sup>*

<sup>1</sup>Communication University of China

<sup>2</sup>Beijing Vocational College of Electronic Science and Technology libobbi@126.com

As a new app, mobile video has attracted a large number of young people with purchasing desire and ability. Taking beauty products as an example, this paper first explores the influencing factors of consumers' purchase intention on the short video platform represented by Tiktok; and builds a model based on SOR theory, which includes consumer stimulus (online sales level of functional and symbolic goods), consumer psychology (consumer identity), consumer reaction (consumer purchase intention), and moderator variables (host characteristics). It is verified that both functional goods with high sales levels and symbolic goods with low sales levels have a positive impact on consumers' purchase intention through the mediating variable (consumer identity). host characteristics play a positive moderating role in the model. Next, this study conducted an in-depth excavation of the moderator variables (characteristics of the hosts). Statistical analysis shows that the beauty industry presents the characteristics of de-gendering of beauty hosts; decentralization of fans and verticalization of anchors. Finally, through the establishment of the generalized additive model, it is known that among the indicators related to the characteristics of hosts, the most influential one is the number of online viewers, and gender has no significant effect on sales.

**Limit Behaviours for Nonlinear Regression Models**

*Yu Miao*

Henan Normal University  
yumiao728@gmail.com

In the report, we shall introduce some recent results for LS and Bayesian estimators in nonlinear regression models, such as, the consistency and deviation inequalities of these estimators.

**Minimum Profile Hellinger Distance Estimation of Covariate Models**

♦*Bowei Ding<sup>1</sup>, Rohana J. Karunamuni<sup>2</sup> and Jingjing Wu<sup>1</sup>*

<sup>1</sup>Department of Mathematics and Statistics, University of Calgary

<sup>2</sup>Department of Mathematical and Statistical Sciences, University of Alberta  
bowei.ding1@ucalgary.ca

Covariate models, such as linear models, generalized linear models and single-index models, are widely used in statistical applications. The importance of such models in statistical analysis is abundantly clear by ever-increasing rate at which articles on covariate models appear in statistical literature. Because of their flexibility, covariate models are being increasingly exploited as a convenient and semi-parametric way to model data that consists of a response variable and covariate variables that affect the response variable. This study investigates efficient and robust estimations for both general covariate models and single-index models, as the most commonly used covariate models in statistical applications. For this purpose, we employ the minimum distance approach. In particular, the minimum Hellinger distance approach introduced by Beran (1977) produces estimators that are asymptotically efficient at the model density and simultaneously possess excellent robustness properties. In this study, we construct several minimum profile Hellinger distance estimators for covariate models and single-index models. We in-

**Session 23CHI39: Recent Advances on Covariate Models and Applications****Low-Tubal-Rank Tensor Sensing and Robust Pca from Quantized Measurements**

*Jianjun Wang*

wjj@swu.edu.cn

Low-rank tensor Sensing (LRTS) is a natural extension of low-rank matrix Sensing (LRMS) to high-dimensional arrays, which aims to reconstruct an underlying tensor  $X$  from incomplete linear measurements  $M(X)$ . However, LRTS ignores the error caused by quantization, limiting its application when the quantization is low-level. Under the tensor Singular Value Decomposition (t-SVD) framework, two recovery methods are proposed. These methods can recover a real tensor  $X$  with tubal rank  $r$  from  $m$  random Gaussian binary measurements with errors decaying at a polynomial speed of the oversampling factor. To improve the convergence rate, we develop a new quantization scheme under which the convergence rate can be accelerated to an exponential function of lambda. Quantized Tensor Robust Principal Component Analysis (Q-TRPCA) aims to recover a low-rank tensor and a sparse tensor from noisy, quantized, and sparsely corrupted measurements. A nonconvex constrained maximum likelihood (ML) estimation method is proposed for Q-TRPCA. We provide an upper bound on the Frobenius norm of tensor estimation error under this method. Making use of tools in information theory, we derive a theoretical lower bound on the best achievable estimation error from unquantized measurements. Com-

## Abstracts

---

vestigate the asymptotic properties, such as uniqueness and consistency, of the proposed estimators. To ease the calculation, a computational algorithm is also developed. Finite-sample performance regarding both efficiency and robustness of the proposed estimators are examined using Monte Carlo simulation studies and real data analysis.

### Session 23CHI46: Advances in Network and Complex Data Analysis

#### Sparse Gaussian Network Model for Single-Cell Rna-Seq Data

*Mingjin Liu<sup>1</sup>, Susmita Datta<sup>1</sup> and ♦Yang Yang<sup>2</sup>*

<sup>1</sup>University of Florida

<sup>2</sup>University of Georgia  
yang.yang4@uga.edu

The single-cell RNA-sequencing (ScRNA) technology has made it possible to trace cellular lineages during differentiation and to identify new cell types. One critical question is whether we can discover the genetic network that controls cellular differentiation and drive transitions from one cell type to another. Unlike bulk transcriptomic data, which average over all cells, scRNA expression data do not obscure biological signals. But such data also pose significant difficulties, for example, substantial cell-to-cell variation and high sparsity caused by dropouts and cell-cycle related effects. Here, we introduce a Bayesian sparse network model based on multivariate normal distribution adapted to handle zero-inflation. This model is designed for cells with a natural temporal order (pseudo-time), e.g., sampled during different stages of cell differentiation. We use cubic B-spline functions to model the mean expression levels of the genes across cells. We then use the Kronecker product between cell-level and gene-level precision matrices to account for the inter-cellular correlation.

#### Nonparametric Regression for 3d Point Cloud Learning

*♦Xinyi Li<sup>1</sup>, Shan Yu, Yueying Wang, Guannan Wang, Li Wang and Ming-Jun Lai*

<sup>1</sup>Clemson University  
lixinyi@clemson.edu

In recent years, there has been an exponentially increased amount of point clouds collected with irregular shapes in various areas. Motivated by the importance of solid modeling for point clouds, we develop a novel and efficient smoothing tool based on multivariate splines over the triangulation to extract the underlying signal and build up a 3D solid model from the point cloud. The proposed method can denoise or deblur the point cloud effectively, provide a multi-resolution reconstruction of the actual signal, and handle sparse and irregularly distributed point clouds to recover the underlying trajectory. In addition, our method provides a natural way of numerosity data reduction. We also introduce a bootstrap method to quantify the uncertainty of the estimators. We establish the theoretical guarantees of the proposed method, including the convergence rate and asymptotic normality of the estimator, and show that the convergence rate achieves optimal nonparametric convergence. Through extensive simulation studies and a real data example, we demonstrate the superiority of the proposed method over traditional smoothing methods in terms of estimation accuracy and efficiency of data reduction. For example, we can reduce an image point cloud of 510,340 voxels to a vector of 4,856 spline coefficients with a peak

#### Nonparametric Link Prediction for Networks and Bipartite Graph

*Jiashen Lu and ♦Kehui Chen*

University of Pittsburgh, USA  
chenkehui@gmail.com

In this talk will discuss a nonparametric link prediction framework for networks and Bipartite graph. In particular, we will discuss how to understand the missing mechanism and to deal with missing observations, when and how to use side information for link prediction, and how to improve the prediction accuracy for new entries (nodes). The proposed statistical framework leads to a simple algorithm with competitive performance.

### Session 23CHI109: Survival Analysis

#### A Weighted Generalized Win-Odds Regression Model for Composite Endpoints

*Bang Wang and ♦Yu Cheng*

University of Pittsburgh  
yucheng95@gmail.com

Time-to-first-event analysis of a composite endpoint is often used for studies involving multiple outcomes. Each component outcome is treated equally, even though they may be of different clinical importance. Win Ratio, net benefit, and Win Odds (WO) have been used as alternative summaries that can handle different types of outcomes and allow for a hierarchical ordering in component outcomes, and thus have drawn much attention in the pharmaceutical industry, academia, and regulatory agencies (Pocock et al., 2012). Most existing work has focused on the nonparametric estimation of these measures under the two-sample setting. In this talk, we propose a proportional win odds regression model to evaluate the treatment effect on multiple outcomes while controlling for other risk factors. The model is easily interpretable as a standard logistic regression model. However, the proposed win odds regression is much more advanced in that multiple numbers and types of outcomes are modeled together and the estimating equation is constructed based on all possible and potentially dependent pairings of a treated subject and a control subject. We also carefully distinguish the ties caused by binary outcomes, which imply treatment equivalency and motivate the use of WO, and the ties due to censored observations. The

#### Kernel Meets Sieve: Transformed Hazards Models with Sparse Longitudinal Covariates

*Dayu Sun<sup>1</sup>, Zhuowei Sun<sup>2</sup>, Xinqiu Zhao<sup>3</sup> and ♦Hongyuan Cao<sup>4</sup>*

<sup>1</sup>Emory University

<sup>2</sup>Jilin University

<sup>3</sup>Hong Kong Polytechnic University

<sup>4</sup>Florida State University

hongyuancao@gmail.com

We study the transformed hazards model with intermittently observed time-dependent covariates for the censored outcome. Existing work assumes the availability of the whole trajectory of the time-dependent covariates, which is not realistic. We propose to combine kernel-weighted log-likelihood and sieve maximum log-likelihood estimation to conduct statistical inference. The method is robust and easy to implement. We establish the asymptotic properties of the proposed estimator and contribute to a rigorous theoretical framework for general kernel-weighted semiparametric M-estimators. Numerical studies corroborate our theoretical results and show that the proposed method has favorable performance over existing methods. An application to a recent COVID-19 study in Wuhan illustrates the practical utility of our method.

**Assessing Dynamic Covariate Effects with Survival Data***Ying Cui and ♦Limin Peng*Emory University  
lpeng@emory.edu

Dynamic (or varying) covariate effects often manifest meaningful physiological mechanisms underlying chronic diseases. However, a static view of covariate effects is typically adopted by standard approaches to evaluating disease prognostic factors, which can result in depreciation of some important disease markers. To address this issue, in this work, we take the perspective of globally concerned quantile regression, and propose a flexible testing framework suited to assess either constant or dynamic covariate effects. We study the powerful Kolmogorov-Smirnov (K-S) and Cramér-Von Mises (C-V) type test statistics and develop a simple resampling procedure to tackle their complicated limit distributions. We provide rigorous theoretical results, including the limit null distributions and consistency under a general class of alternative hypotheses of the proposed tests, as well as the justifications for the presented resampling procedure. Extensive simulation studies and a real data example demonstrate the utility of the new testing procedures and their advantages over existing approaches in assessing dynamic covariate effects.

**Doubly Robust Estimation under Covariate-Induced Dependent Left Truncation***Yuyao Wang<sup>1</sup>, Andrew Ying<sup>2</sup> and ♦Ronghui Xu<sup>1</sup>*<sup>1</sup>University of California, San Diego<sup>2</sup>Google Inc.  
rxu@ucsd.edu

In prevalent cohort studies with follow-up, the time-to-event outcome is subject to left truncation leading to selection bias. For estimation of the distribution of time-to-event, conventional methods adjusting for left truncation tend to rely on the (quasi-)independence assumption that the truncation time and the event time are “independent” on the observed region. This assumption is violated when there is dependence between the truncation time and the event time possibly induced by measured covariates. Inverse probability of truncation weighting leveraging covariate information can be used in this case, but it is sensitive to misspecification of the truncation model. In this work, we apply the semiparametric theory to find the efficient influence curve of an expected (arbitrarily transformed) survival time in the presence of covariate-induced dependent left truncation. We then use it to construct estimators that are shown to enjoy double-robustness properties. Our work represents the first attempt to construct doubly robust estimators in the presence of left truncation, which does not fall under the established framework of coarsened data where doubly robust approaches are developed. We provide technical conditions for the asymptotic properties that appear to not have been carefully examined in the literature for time-to-event data

**Session 23CHI120: Advance in Statistical Methods for Large and Complex Data****Nonlinear Expectation for Bandit Learning***Xiaodong Yan*Shandong University  
yanxiaodong128@163.com

Recently, the Nonlinear Expectation Team at Shandong University has developed a series of limit theory based on multi-armed bandits, which are important research achievements in the field of Nonlinear

Expectation Theory. This report considers the statistical issues related to this original theoretical achievement and proposes new statistical machine learning methods, including (1) optimal distribution and paradox in the context of two-armed bandits, (2) two-sample testing for big data, (3) proposing a sequential bootstrap method for hypothesis testing and a general framework for single-sample testing, (4) finally, to consider generalized models, we propose a parameter testing method based on stochastic gradient descent. It is worth noting that the new statistical machine learning methods fully integrate knowledge, which belongs to the statistical learning methods combining knowledge reasoning and data-driven approaches. Moreover, the new methods for hypothesis testing have achieved higher power compared to traditional methods.

**Functional Linear Operator Quantile Regression for Sparse Longitudinal Data***♦Xingcai Zhou<sup>1</sup>, Tingyu Lai<sup>2</sup> and Linglong Kong<sup>3</sup>*<sup>1</sup>Nanjing Audit University<sup>2</sup>Guangxi Normal University<sup>3</sup>University of Alberta  
xzhou@nuaa.edu.cn

We propose a functional linear operator quantile regression (FLOQR) framework, which includes many important and useful functional data models, and devote to the new framework model for longitudinal data with the typically sparse and irregular designs. The non-smooth quantile loss and functional linear operator pose new challenges to functional data analysis for longitudinal data in both computation and theoretical development. To address the challenge, we propose the iterative surrogate least squares estimation approach for the FLOQR model, which transforms the response trajectories and establishes a new connection between FLOQR and functional linear operator model. In addition, we use Karhunen-Loeve expansion to alleviate the problem of the nonexistence of the inverse of the covariance in the infinite dimensional Hilbert space. Then, the approach is used to classic functional varying coefficient QR, functional linear QR and functional varying coefficient QR with history index function for sparse longitudinal data by using functional principal components analysis through conditional expectation. The resulting technique is flexible and allows the prediction of an unobserved quantile response trajectory from sparse measurements of a predictor trajectory. Theoretically, we show that, after a constant number of iterations, the proposed estimator is asymptotic consistent for sparse designs. Moreover,

**Order Statistics Approaches to Unobserved Heterogeneity in Auctions***Yao Luo<sup>1</sup>, ♦Peijun Sang<sup>2</sup> and Ruli Xiao<sup>3</sup>*<sup>1</sup>University of Toronto<sup>2</sup>University of Waterloo<sup>3</sup>Indiana University  
psang@uwaterloo.ca

We establish nonparametric identification of auction models with continuous and nonseparable unobserved heterogeneity using three consecutive order statistics of bids. We then propose sieve maximum likelihood estimators for the joint distribution of unobserved heterogeneity and the private value, as well as their conditional and marginal distributions. Lastly, we apply our methodology to a novel dataset from judicial auctions in China. Our estimates suggest substantial gains from accounting for unobserved heterogeneity when setting reserve prices. We propose a simple scheme that achieves nearly optimal revenue by using the appraisal value as the reserve price.

## Abstracts

---

### Significance Test for Multinomial Naive Bayes Classifier with Ultra-High Dimensional Binary Features

♦*Baiguo An, Juan Zhang, Beibei Zhang and Wenliang Pan*  
anbg200@163.com

We developed a significance test method for multinomial naive bayes classifier with ultra-high dimensional binary features. A novel test statistic with asymptotic standard Gaussian null distribution is proposed. Under very mild assumptions, the proposed test statistic has powers that tend to 1 as the sample size tends to infinity. Then a sequential test process is developed to perform variable screening. We applied the proposed methods to lots of numerical studies including simulated examples and two real text data classification examples. The results show that our methods have good finite sample performances.

### Session 23CHI121: Modern Statistical and Machine Learning Methods for Analysis of -Omics Data

#### Integrative Analysis of Gene Expression and Histology Images in Spatial Transcriptomics

♦*Mingyao Li and Daiwei Zhang*

University of Pennsylvania  
mingyao@pennmedicine.upenn.edu

Recent developments in spatial transcriptomics technologies have enabled scientists to get an integrated understanding of cells in their morphological context. Applications of these technologies in diverse tissues and diseases have transformed our views of transcriptional complexity. Most published studies utilized tools developed for single-cell RNA-seq for data analysis. However, spatial transcriptomics data exhibit different properties from single-cell RNA-seq. To take full advantage of the added dimension in spatial location information in such data, new methods that are tailored for spatial transcriptomics are needed. Additionally, spatial transcriptomics data often have companion high-resolution histology information available. Incorporating histological features in gene expression analysis is an underexplored area. In this talk, I will discuss how to best extract features from histology images and how to integrate these features with gene expression to detect spatial domains and automatically annotate tissue sections in spatial transcriptomics. I will show applications of these methods in spatial transcriptomics data generated from brain and cancer tissues.

#### Bayesian Pathway Analysis over Brain Network Mediators and Genetic Exposure for Survival Data

*Xinyuan Tian<sup>1</sup>, Fan Li<sup>1</sup>, Li Shen<sup>2</sup>, Denise Esserman<sup>1</sup> and ♦Yize Zhao<sup>1</sup>*

<sup>1</sup>Yale University

<sup>2</sup>University of Pennsylvania  
yize.zhao@yale.edu

Technological advancements in noninvasive imaging facilitate the construction of whole brain inter-connected networks, known as brain connectivity. Existing approaches to analyze brain connectivity frequently disaggregate the entire network into a vector of unique edges or summary measures, leading to a substantial loss of information. Motivated by the need to explore the effect mechanism among genetic exposure, brain connectivity and time to disease onset, we propose an integrative Bayesian framework to model the effect pathway between each of these components while quantifying the mediating role of brain networks. To accommodate the biological architectures of brain connectivity constructed along white matter fiber tracts, we develop a structural modeling framework which

includes a symmetric matrix-variate accelerated failure time model and a symmetric matrix response regression to characterize the effect paths. We further impose within-graph sparsity and between-graph shrinkage to identify informative network configurations and eliminate the interference of noisy components. Extensive simulations confirm the superiority of our method compared with existing alternatives. By applying the proposed method to the landmark Alzheimer's Disease Neuroimaging Initiative study, we obtain neurobiologically plausible insights that may inform future intervention strategies.

#### Synthetic Rna Sequencing Data from Pilot Studies using Deep Generative Models

♦*Yunhui Qi<sup>1</sup>, Xinyi Wang<sup>2</sup> and Lixuan Qin<sup>3</sup>*

<sup>1</sup>Iowa State University

<sup>2</sup>Columbia University

<sup>3</sup>Memorial Sloan Kettering Cancer Center  
qyh@iastate.edu

Deep generative models such as variational auto-encoder, generative adversarial network and flow-based generative models are able to learn the distribution of data and generate a large number of samples from the learned distribution. They have been successfully applied to augment the imaging data, text data and even single cell RNA sequencing data. The success of these applications relies on large training sample size and high signal to noise ratio. The performance of deep generative models on bulk RNA-seq data which has relatively small sample size and signal to noise ratio is unclear. In this paper, we evaluate the performance of three types of deep generative models using TCGA microRNA and RNA sequencing data under a variety of conditions including sample sizes, expression level, noise level and multiple tuning parameters. We show that deep generative models can work on bulk RNA-seq data under certain circumstances and give suggestions on how to use them.

#### Differential Inference for Single-Cell Rna-Sequencing Data

♦*Fangda Song<sup>1</sup>, Kelvin Y. Yip<sup>2</sup> and Yingying Wei<sup>2</sup>*

<sup>1</sup>The Chinese University of Hong Kong, Shenzhen

<sup>2</sup>The Chinese University of Hong Kong  
songfangda@cuhk.edu.cn

With the wide adoption of single-cell RNA-seq (scRNA-seq) technologies, scRNA-seq experiments are becoming more and more complicated with multiple treatments or biological conditions. However, despite the active research on batch effects correction, cell type clustering, and missing data imputation for scRNA-seq data, rigorous statistical methods to compare scRNA-seq experiments under different conditions are still lacking. Here, we propose a Bayesian hierarchical model, DIFFerential Inference for Single-cell RNA-sequencing Data (DIFseq), to rigorously quantify the treatment effects on both cellular compositions and cell-type-specific gene expression levels for scRNA-seq data. We derive conditions for the model identifiability, which provides guidelines on the experimental design for comparative scRNA-seq studies. We implement a highly scalable Monte Carlo Expectation-Maximization algorithm to handle a large number of cells. Application of DIFseq to a pancreatic study demonstrates that considering the biological conditions of samples in the analysis substantially boosts the clustering accuracy as compared to traditional analysis pipelines for scRNA-seq data and identifies cell-type-specific and condition-specific differentially expressed genes.

## Session 23CHI125: Large Dimensional Random Matrix and Its Applications

### Block-Diagonal Test for High-Dimensional Covariance Matrices

Jiayu Lai<sup>1</sup>, ♦Xiaoyi Wang<sup>2</sup>, Shurong Zheng<sup>1</sup> and Kaige Zhao<sup>1</sup>

<sup>1</sup>Northeast Normal University

<sup>2</sup>Beijing Normal University  
wangxy059@nenu.edu.cn

The testing structure of a high-dimensional covariance matrix plays an important role in financial stock analyses, genetic series analyses, and many other fields. Testing that the covariance matrix is block-diagonal under the high-dimensional setting is a main focus of this paper. To tackle this problem, test procedures that are powerful under normality assumptions, two-diagonal block assumptions or sub-block dimensionality assumptions have been proposed in several existing studies. To relax these conditions, a test framework based on U-statistics is proposed in this paper, and the asymptotic distributions of those U-statistics are established under the null and alternative hypotheses. Moreover, another test approach is developed for alternatives with different sparsity levels. Finally, both a simulation study and real data analysis are conducted to show the performance of our proposed test procedures.

### Estimating the Number of Communities Based on Individual-Centered Partial Information

Xiyue Zhu, Xiao Han and ♦Qing Yang

University of Science and Technology of China  
yangq@ustc.edu.cn

Considering a new community detection method based on the partial network information, we establish a test hypothesis to estimate the number of communities K adaptively from data. Sum of the singular values and eigenvalues of partitioned matrices obtained from a centered and rescaled partial adjacency matrix constitute the test statistic, which converges to the sum of two Tracy-Widom distributions. We derive the asymptotic null distribution and consistency of estimating K by a series of important results in random matrix theory. The validity of the model is also performed in extensive simulations and real data examples, including both directed graphs and undirected graphs.

### On the Asymptotic Properties of Spike Eigenvalues and Eigenvectors of Signal-Plus-Noise Matrices with their Applications

♦Yiming Liu<sup>1</sup>, Zhixiang Zhang<sup>2</sup>, Guangming Pan<sup>3</sup> and Lingyue Zhang

<sup>1</sup>Jinan University

<sup>2</sup>University of Pennsylvania

<sup>3</sup>NANYANG TECHNOLOGICAL UNIVERSITY  
liuyiming@jnu.edu.cn

This paper is to investigate the asymptotics properties of the spike eigenvalues and the corresponding eigenvectors under a general low rank signal plus noise model in high dimension. Under mild conditions with respect to the leading eigenvalue of the underlying covariance matrix and the noises, we find the limits of both spike eigenvalues and eigenvectors of the sample covariance matrix. Based on the discovered results, some related applications are also considered. Specifically, for a general mixture model, a new criterion to estimate the number of clusters is proposed; the properties of spectral clustering is also investigated. In addition, some classification and dimension reduction problems are also considered.

### Random Matrix Methods for Machine Learning: An Applica-

### tion to “lossless” Compression of Deep Neural Networks

Zhenyu Liao

Huazhong University of Science and Technology (HUST)  
zhenyu\_liao@hust.edu.cn

The advent of the Big Data era has triggered a renewed interest in large-dimensional machine learning (ML) and in particular, deep neural networks (DNNs). These methods, being developed from small-dimensional intuitions, often behave dramatically differently from their original designs and tend to be inefficient on large-dimensional datasets. Leveraging the fact that both the dimension and the size of datasets are large, recent advances in random matrix theory (RMT) provide novel insights, allowing for a renewed understanding and the possibility to design more efficient machine learning approaches, thereby opening the door to completely new paradigms. In this talk, we will discuss the curse of dimensionality and concentration phenomenon in high dimensions, and highlight how these counterintuitive phenomena arise in ML practice when large-dimensional data are considered. By focusing on the urgent need of compressing large and deep neural networks, we show how the proposed random matrix methods can be applied to design efficient DNN compression schemes with strong performance guarantees.

## Session 23CHI126: Recent Developments in Biostatistics and Beyond

### STAARpipeline: An all-in-One Rare-Variant Tool for Biobank-Scale Whole-Genome Sequencing Data

♦Zilin Li<sup>1</sup>, Xihao Li<sup>2</sup> and Xihong Lin<sup>2</sup>

<sup>1</sup>Indiana University School of Medicine

<sup>2</sup>Harvard T.H. Chan School of Public Health  
muzizimumu@gmail.com

Large-scale whole-genome sequencing (WGS) studies have enabled the analysis of rare variant associations with complex human diseases and traits. Variant set analysis is a powerful approach to studying rare variant associations. However, existing methods have limited ability to define the variant set in the genome, especially for the noncoding genome. We propose a computationally efficient and robust rare variant association-detection framework, STAARpipeline, to automatically annotate a WGS study and perform flexible rare variant association analysis, including gene-centric analysis and fixed-window and dynamic-window-based non-gene-centric analysis by incorporating variant functional annotations. In gene-centric analysis, STAARpipeline groups coding and noncoding variants based on functional categories of genes and incorporate multiple functional annotations. In non-gene-centric analysis, in addition to fixed-size sliding window analysis, STAARpipeline provides a data-adaptive-size dynamic window analysis. All these variant sets could be automatically defined and selected in STAARpipeline. STAARpipeline also provides analytical follow-up of dissecting association signals independent of known variants via conditional analysis. We applied the STAARpipeline to analyze the total cholesterol in 30,138 samples from the NHLBI Trans-Omics for Precision Medicine (TOPMed) Program. All analyses scale well in computation time and memory. We discover several potentially new significant associations with lipids, including a finding of rare variants in

### Bayesian Sparse Gaussian Mixture Model for Clustering in High Dimensions

♦Dapeng Yao<sup>1</sup>, Fangzheng Xie<sup>2</sup> and Yanxun Xu<sup>1</sup>

## Abstracts

---

<sup>1</sup>Johns Hopkins University

<sup>2</sup>Indiana University

dyao10@jhu.edu

We study the sparse high-dimensional Gaussian mixture model when the number of clusters is allowed to grow with the sample size. A minimax lower bound for parameter estimation is established, and we show that a constrained maximum likelihood estimator achieves the minimax lower bound. However, this optimization-based estimator is computationally intractable because the objective function is highly nonconvex and the feasible set involves discrete structures. To address the computational challenge, we propose a computationally tractable Bayesian approach to estimate high-dimensional Gaussian mixtures whose cluster centers exhibit sparsity using a continuous spike-and-slab prior. We further prove that the posterior contraction rate of the proposed Bayesian method is minimax optimal. The mis-clustering rate is obtained as a by-product using tools from matrix perturbation theory. The proposed Bayesian sparse Gaussian mixture model does not require pre-specifying the number of clusters, which can be adaptively estimated. The validity and usefulness of the proposed method is demonstrated through simulation studies and the analysis of a real-world single-cell RNA sequencing data set.

### Prediction-Assisted Candidate Screening with Fdr Control via Conformal Inference

♦Ying Jin and Emmanuel Candès

Stanford University

ying531@stanford.edu

Decision making or scientific discovery pipelines such as job hiring and drug discovery often involve multiple stages: before any resource-intensive step, there is often an initial screening that uses predictions from a machine learning model to shortlist a few candidates from a large pool. We study screening procedures that aim to select candidates whose unobserved outcomes exceed user-specified values. We develop a method that wraps around any prediction model to produce a subset of candidates while controlling the proportion of falsely selected units. Building upon the conformal inference framework, our method first constructs p-values that quantify the statistical evidence for large outcomes; it then determines the shortlist by comparing the p-values to a threshold introduced in the multiple testing literature. In many cases, the procedure selects candidates whose predictions are above a data-dependent threshold. Our theoretical guarantee holds under mild exchangeability conditions on the samples, generalizing existing results on multiple conformal p-values. We demonstrate the empirical performance of our method via simulations, and apply it to job hiring and drug discovery datasets.

### Session 23CHI127: Experimental Designs and their Applications (II)

#### Statistical Inference Through Combining Physical and Computer Experiments

Yang Li and ♦Shifeng Xiong

xiong@amss.ac.cn

Computer simulation models are widely used to study complex physical systems. A related fundamental topic is the inverse problem, also called calibration, which aims at estimating parameters in the model based on observations. We propose a semi-parametric model, called the discrepancy decomposition model, to describe the discrepancy between the physical system and the computer model.

The proposed model possesses a clear interpretation, and more importantly, it is identifiable under mild conditions. Under this model, we present estimators of the parameters and the discrepancy, and then establish their asymptotic properties. Hypothesis testing problems are also discussed with both physical observations and computer experiments. Numerical examples show the effectiveness of the proposed methods.

#### A Sequential Design for Determining the Quantile Curve in Two-Factor Sensitivity Tests

Yuxia Liu, ♦Yubin Tian and Dianpeng Wang

Beijing Institute of Technology

tianyb@bit.edu.cn

Many works in the literature have proposed methods for estimating quantiles in sensitivity experiments with a single factor. However, sensitivity experiments with multiple factors are more and more popular in practice. A significant important challenge remains the estimation of quantiles of sensitivity experiments with multiple factors, which are curves or surfaces. While Some methods have been proposed to estimate the parameters in such models, the estimation of quantiles has received scant attention. Motivated by this problem and applications for field use, this study proposes a new three-phase optimal method for sensitivity experiments with two factors to estimate the interested quantiles directly. The simulation studies and the illustrative application to a pyrotechnic control subsystem demonstrate that the proposed method has excellent performance.

#### A Subsampling Method for Regression Problems Based on Minimum Energy Criterion

Wenlin Dai<sup>1</sup>, Yan Song<sup>1</sup> and ♦Dianpeng Wang<sup>2</sup>

<sup>1</sup>Renmin University of China

<sup>2</sup>Beijing institute of technology

wdp@bit.edu.cn

The extraordinary amounts of data generated nowadays pose heavy demands on computational resources and time, which hinders the implementation of various statistical methods. An efficient and popular strategy of downsizing data volumes and thus alleviating these challenges is subsampling. However, the existing methods either rely on specific assumptions for the underlying models or acquire partial information from the available data. For regression problems, we propose a novel approach, termed adaptive subsampling with the minimum energy criterion (ASMEC). The proposed method requires no explicit model assumptions and “smartly” incorporates information on covariates and responses. ASMEC subsamples possess two desirable properties: space-fillingness and spatial adaptiveness. We investigate the limiting distribution of ASMEC subsamples and their theoretical properties under the smoothing spline regression model. The effectiveness and robustness of the ASMEC approach are also supported by a variety of synthetic examples and two real-life examples.

#### Global Likelihood Sampler for Multimodal Distributions

♦Si-Yu Yi, Ze Liu, Min-Qian Liu and Yong-Dao Zhou

Nankai University

siyuyi@mail.nankai.edu.cn

Drawing samples from a target distribution is essential for statistical computations when the analytical solution is infeasible. Many existing sampling methods may be easy to fall into the local mode or strongly depend on the proposal distribution when the target distribution is complicated. In this talk, we introduce our proposed Global Likelihood Sampler (GLS) to tackle these problems and use the GL bootstrap to assess the Monte Carlo error. GLS takes the advantage of the randomly shifted low-discrepancy point set to suf-

ficiently explore the structure of the target distribution. It is efficient for multimodal and high-dimensional distributions and easy to implement. It is shown that the empirical cumulative distribution function of the samples uniformly converges to the target distribution under some conditions. The convergence for the approximate sampling distribution of the sample mean based on the GL bootstrap is also obtained. Moreover, numerical experiments and a real application are conducted to show the effectiveness, robustness, and speediness of GLS compared with some common methods. It illustrates that GLS can be a competitive alternative to existing sampling methods.

## Session 23CHI128: Space-Filling Designs (II)

### Efficient Kriging using Designs with Low Fill Distance and High Separation Distance

Xu He

Chinese Academy of Sciences  
hexu@amss.ac.cn

Kriging is a powerful technique to emulate computer experiments. Accurate Kriging requires good experimental designs. The purpose of this paper is to construct space-filling designs that are most suitable for Kriging. Firstly, we provide evidences that although it is well-known that designs with low fill distance or high separation distance are appealing for Kriging, the two distance measures should be combined to reach a more striking criterion for selecting designs. Secondly, we propose a method to efficiently construct optimal designs under the newly proposed criterion. Thirdly, we provide a robust method to transform design points towards boundary facets of the input space. Numerical results suggest that the new criterion, construction, and transformation combined perform well in Kriging interpolation under various scenarios.

### Doubly Coupled Designs for Computer Experiments with both Qualitative and Quantitative Factors

♦Feng Yang<sup>1</sup>, C. Devon Lin<sup>2</sup>, Yongdao Zhou<sup>3</sup> and Yuanzhen He<sup>4</sup>

<sup>1</sup>Sichuan Normal University

<sup>2</sup>Queen's University

<sup>3</sup>Nankai University

<sup>4</sup>Beijing Normal University  
yangfeng@sicnu.edu.cn

Computer experiments with both qualitative and quantitative input variables occur frequently in many scientific applications. As a result, how to choose the input settings for such experiments is important. SLHDs were the first systematic approach to address this issue. However, the cost of such designs increase with an increasing number of level combinations of the qualitative factors. To reduce the cost, MCDs were proposed where the design for the quantitative factors is a SLHD with respect to each qualitative factor. The drawback of such designs is that the corresponding data may not be able to capture the effects between any two (and more) qualitative factors and quantitative factors. To balance the run size and design efficiency, we propose a new type of designs, doubly coupled designs, where the design points for the quantitative factors form a SLHD with respect to the levels of any qualitative factor and with respect to the level combinations of any two qualitative factors, respectively. The proposed designs have a better stratification property compared with that of MCD. @e establish the existence of the proposed designs, introduce several construction methods and examine the properties of the resulting designs.

### Deterministic Construction Methods for Uniform Designs

♦Liang-Wei Qi, Ze Liu and Yong-Dao Zhou

School of Statistics and Data Science, Nankai University  
lwqi@mail.nankai.edu.cn

Space-filling designs are useful for exploring the relationship between the response and factors, especially when the true model is unknown. The wrap-around L2-discrepancy is an important measure of the uniformity, and has often been used as a type of space-filling criterion. However, most obtained designs are generated through stochastic optimization algorithms, and cannot achieve the lower bound of the discrepancies and are only nearly uniform. Then deterministic construction methods for uniform designs are desired. This paper constructs uniform designs under the wrap-around L2-discrepancy by generator matrices of linear codes. Several requirements on the generator matrices, such as a necessary and sufficient condition for generating uniform designs, are derived. Based on these, two simple deterministic constructions for uniform designs are given. Some examples illustrate the effectiveness of them. Moreover, the resulting designs can be regarded as a generalization of good lattice point sets, and also enjoy good orthogonality.

### Deterministic Construction Methods for Uniform Designs

♦Liangwei Qi, Ze Liu and Yongdao Zhou

Nankai University  
ydzhou@nankai.edu.cn

Space-filling designs are useful for exploring the relationship between the response and factors, especially when the true model is unknown. The wrap-around L2-discrepancy is an important measure of the uniformity, and has often been used as a type of space-filling criterion. However, most obtained designs are generated through stochastic optimization algorithms, and cannot achieve the lower bound of the discrepancies and are only nearly uniform. Then deterministic construction methods for uniform designs are desired. This paper constructs uniform designs under the wrap-around L2-discrepancy by generator matrices of linear codes. Several requirements on the generator matrices, such as a necessary and sufficient condition for generating uniform designs, are derived. Based on these, two simple deterministic constructions for uniform designs are given. Some examples illustrate the effectiveness of them. Moreover, the resulting designs can be regarded as a generalization of good lattice point sets, and also enjoy good orthogonality.

### A Minimum Aberration Type Criterion for Selecting Space-Filling Designs.

♦Ye Tian<sup>1</sup> and Hongquan Xu<sup>2</sup>

<sup>1</sup>Beijing University of Posts and Telecommunications

<sup>2</sup>University of California, Los Angeles  
yetianhuawu@hotmail.com

Space-filling designs are widely used in computer experiments. Inspired by the stratified orthogonality of strong orthogonal arrays, we propose a minimum aberration type criterion for assessing the space-filling properties of designs based on design stratification properties on various grids. A space-filling hierarchy principle is proposed as a basic assumption of the criterion. The new criterion provides a systematic way of classifying and ranking space-filling designs including various types of strong orthogonal arrays and Latin hypercube designs. Theoretical results and examples are presented to show that strong orthogonal arrays of maximum strength are favorable under the proposed criterion. For strong orthogonal arrays with the same strength, the space-filling criterion can further rank them based on their space-filling patterns.

## Abstracts

---

### Session 23CHI133: Statistical and Machine Learning Methods for Biomedical Research

#### Fusion of Supervised Learning and Reinforcement Learning for Dynamic Treatment Recommendation

Yiyuan Liu<sup>1</sup>, Linjiajie Fang<sup>2</sup>, Qiyue Wang<sup>2</sup> and ♦Bingyi Jing<sup>3</sup>

<sup>1</sup>Jiangxi Univ. of Finance & Economics

<sup>2</sup>HKUST

<sup>3</sup>SUSTech

jingby@sustech.edu.cn

Electronic health records (EHR) have provided a great opportunity to exploit personalized health data to optimize clinical decision making and achieve personalized treatment recommendation. In this talk, we explore how AI could help physicians in prescribing medicines for patients with multi-morbidity (i.e., co-occurrence of two or more diseases). Both Supervised Learning (SL) and Reinforcement Learning (RL) have been employed for this purpose, but with their own drawbacks. For instance, SL relies highly on the clinical guideline and doctors personal experience while RL may produce unacceptable medications due to lack of the supervision from doctors. In this talk, we propose a novel SAVER framework by fusing RL and SL, where RL learns the optimal policy and SL gives a regularization to avoid unacceptable risks. Our experiments show that our SAVER framework can provider more accuracy treatment recommendation than the existing methods.

#### Deep Learning for Cell Type Identification Base on Single-Cell Chromatin Accessibility Data

Rui Jiang

清华大学

ruijiang@tsinghua.edu.cn

The human body consists of about 50 trillion cells. The identification of their cell types based on the status of biomolecules is of great importance to not only the understanding of fundamental biological questions such as the development of human body and the differentiation of cells, but also downstream biomedical applications such as the diagnosis of diseases and the development of novel drugs. Targeting on cell type identification based on single-cell chromatin accessibility data, this talk introduces deep learning models designed in unsupervised, weakly supervised, and supervised manner, including the following parts. First, a deep generative neural network called Roundtrip for dealing with high dimensional and sparse single-cell data. Second, an unsupervised deep neural network called scDEC for cell clustering. Third, a weakly supervised model called RA3 for cell clustering with the use of bulk sequencing data as the reference. Finally, a supervised generative model called EpiAnno for cell type annotation based on cells with known cell types. The development of these methods follows the logic of utilizing information of cell types contained in the vast amount of reference genomic sequencing data in a more and more sophisticated manner. It is expected that these methods can greatly improve the identification

#### A Fast Method for Inference of Phylogenetic Networks

Louxin Zhang

National University of Singapore

matzlx@nus.edu.sg

Phylogenetic trees have been used to model the evolution of species for over 200 years. However, phylogenetic networks are more useful than trees for describing and visualizing recombination, hybridization and horizontal gene transfer events. Since the space of phylogenetic networks is much larger than the space of phylogenetic trees, it is extremely challenging to infer phylogenetic net-

works from gene trees and from sequence data. Here, the talk will introduce a recent parsimonious method named ALTS for inferring phylogenetic network networks. It is based on a reduction from the minimum phylogenetic network inference problem to the shortest common super-sequence problem.

#### Calibrating p Values for Peptide Identification

Juntao Zhao<sup>1</sup>, Sheng Lian<sup>1</sup>, Zhen Zhang<sup>2</sup>, Xiaodan Fan<sup>2</sup>, Ning Li<sup>1</sup> and ♦Weichuan Yu<sup>1</sup>

<sup>1</sup>HKUST

<sup>2</sup>CUHK

eeyu@ust.hk

Peptide identification is a fundamental step in the analysis of mass spectrometry data for computational proteomics studies. The reliability of peptide identification results is commonly measured using p value or its variations. However, different peptide identification methods produce different p values and no calibration procedure is available so far in the community of computational proteomics. In this talk, we propose a new procedure to calibrate p values for peptide identification. Experimental results demonstrate that the calibrated p values are more accurate than those provided by existing peptide identification methods.

### Session 23CHI137: Recent Developments in Machine Learning and Causal Inference

#### Ensemble Methods for Testing a Global Null

Yaowu Liu

Southwestern University of Finance and Economics

yaowuliu615@gmail.com

Testing a global null is a canonical problem in statistics and has a wide range of applications. In view of the fact that no uniformly most powerful test exists, prior and/or domain knowledge are commonly used to focus on a certain class of alternatives to improve the testing power. However, it is generally challenging to develop tests that are particularly powerful against a certain class of alternatives. In this paper, motivated by the success of ensemble learning methods for prediction or classification, we propose an ensemble framework for testing that mimics the spirit of random forests to deal with the challenges. Our ensemble testing framework aggregates a collection of weak base tests to form a final ensemble test that maintains strong and robust power. We apply the framework to four problems about global testing in different classes of alternatives arising from Whole Genome Sequencing (WGS) association studies. Specific ensemble tests are proposed for each of these problems, and their theoretical optimality is established in terms of Bahadur efficiency. Extensive simulations and an analysis of a real WGS dataset are conducted to demonstrate the type I error control and/or power gain of the proposed ensemble tests.

#### Policy Learning with Asymmetric Utilities

Eli Ben-Michael<sup>1</sup>, Kosuke Imai<sup>2</sup> and ♦Zhichao Jiang<sup>3</sup>

<sup>1</sup>Carnegie Mellon University

<sup>2</sup>Harvard University

<sup>3</sup>Sun Yat-sen University

jiangzhch7@mail.sysu.edu.cn

Learning optimal policies from observed data requires a careful formulation of the utility function whose expected value is maximized across a population. Although researchers typically use utilities that depend on observed outcomes alone, in many settings the decision maker's utility function is more properly characterized by the joint

set of potential outcomes under all actions. For example, the Hippocratic principle to “do no harm” implies that the cost of causing death to a patient who would otherwise survive without treatment is greater than the cost of forgoing life-saving treatment. We consider optimal policy learning with asymmetric utility functions of this form. We show that asymmetric utilities lead to an unidentifiable social welfare function, and so we first partially identify it. Drawing on statistical decision theory, we then derive minimax decision rules by minimizing the maximum regret relative to alternative policies. We learn minimax decision rules from observed data by solving intermediate classification problems.

### Semiparametric Proximal Causal Inference

♦Yifan Cui<sup>1</sup>, Hongming Pu<sup>2</sup>, Xu Shi<sup>3</sup>, Wang Miao<sup>4</sup> and Eric Tchetgen Tchetgen<sup>2</sup>

<sup>1</sup>ZJU

<sup>2</sup>UPenn

<sup>3</sup>UMich

<sup>4</sup>PKU

yifanc095@gmail.com

Skepticism about the assumption of no unmeasured confounding, also known as exchangeability, is often warranted in making causal inferences from observational data; because exchangeability hinges on an investigator’s ability to accurately measure covariates that capture all potential sources of confounding. In practice, the most one can hope for is that covariate measurements are at best proxies of the true underlying confounding mechanism operating in a given observational study. In this paper, we consider the framework of proximal causal inference introduced by Miao et al. (2018); Tchetgen Tchetgen et al. (2020), which while explicitly acknowledging covariate measurements as imperfect proxies of confounding mechanisms, offers an opportunity to learn about causal effects in settings where exchangeability on the basis of measured covariates fails. We make a number of contributions to proximal inference including (i) an alternative set of conditions for nonparametric proximal identification of the average treatment effect; (ii) general semiparametric theory for proximal estimation of the average treatment effect including efficiency bounds for key semiparametric models of interest; (iii) a characterization of proximal doubly robust and locally efficient estimators of the average treatment effect. Moreover, we provide analogous identification and efficiency results for the average treatment effect on the treated.

### New $\sqrt{n}$ -Consistent, Numerically Stable Higher Order Influence Function Estimators

Lin Liu

Shanghai Jiao Tong University  
linliu.tju@gmail.com

Higher-Order Influence Functions (HOIFs) provide a unified theory for constructing rate-optimal estimators for a large class of low-dimensional (smooth) statistical functionals/parameters (and sometimes even infinite-dimensional functions) that arise in substantive fields including epidemiology, economics and the social sciences. Since the introduction of HOIFs by Robins et al. (2008), they have been viewed mostly as a theoretical benchmark rather than a useful tool for statistical practice. Works aimed to flip the script are scant, but a few recent papers (Liu et al. 2017, 2021b) make some partial progress. In this paper, we take a fresh attempt at achieving this goal by constructing new, numerically stable HOIF estimators (or SHOIF estimators for short with “s” standing for “stable”) with provable statistical, numerical and computational guarantees. This new class of SHOIF estimators (up to the 2nd order) was foreshadowed

in synthetic experiments conducted by Liu et al. (2020a).

### Session 23CHI138: Recent Developments in Applied Probability and Statistics

#### High-Dimensional Dimension Reduction and Its Application to Classification

♦Zhibo Cai<sup>1</sup>, Yingcun Xia<sup>2</sup> and Weiqiang Hang<sup>2</sup>

<sup>1</sup>Renmin University of China

<sup>2</sup>Nantional University of Singapore  
caizhibo@ruc.edu.cn

Sufficient Dimension Reduction (SDR) selects important linear combinations of predictors to reduce the dimension of data. However, its ability to improve general function estimation or classification has not been well received, especially for high-dimensional data. In this talk, we first introduce a local liner smoother for high-dimensional nonparametric regression and then utilize it in the outer-product-of-gradient (OPG) approach of SDR to high-dimensional data, which allows the dimension to diverge to infinity with sample size. We call the method high-dimensional OPG (HOPG) and show that it is consistent. Superior performance over the counterparts of HOPG is observed in simulations. To apply SDR to classification in high-dimensional data, we propose an ensemble classifier by aggregating results of classifiers that are built on subspaces reduced by the random projection and HOPG consecutively from the original data. The real data classification problems are studied compared with some popular classifiers, which shows encouraging results.

#### From p-Wasserstein Bounds to Moderate Deviations

♦Xiao Fang<sup>1</sup> and Yuta Koike<sup>2</sup>

<sup>1</sup>The Chinese University of Hong Kong

<sup>2</sup>University of Tokyo  
xfang@sta.cuhk.edu.hk

We use a new method via  $p$ -Wasserstein bounds to prove Cramér-type moderate deviations in (multivariate) normal approximations. In the classical setting that  $W$  is a standardized sum of  $n$  independent and identically distributed (i.i.d.) random variables with sub-exponential tails, our method recovers the optimal range of  $0 \leq x = o(n^{1/6})$  and the near optimal error rate  $O(1)(1+x)(\log n + x^2)/\sqrt{n}$  for  $P(W > x)/(1 - \Phi(x)) \rightarrow 1$ , where  $\Phi$  is the standard normal distribution function. Our method also works for dependent random variables (vectors) and we give applications to the combinatorial central limit theorem, Wiener chaos, homogeneous sums and local dependence. The key step of our method is to show that the  $p$ -Wasserstein distance between the distribution of the random variable (vector) of interest and a normal distribution grows like  $O(p^\alpha \Delta)$ ,  $1 \leq p \leq p_0$ , for some constants  $\alpha, \Delta$  and  $p_0$ . In the above i.i.d. setting,  $\alpha = 1, \Delta = 1/\sqrt{n}, p_0 = n^{1/3}$ . For this purpose, we obtain general  $p$ -Wasserstein bounds in (multivariate) normal approximations using Stein’s method.

#### A Two-Way Heterogeneity Model for Dynamic Networks

♦Bin Yan Jiang<sup>1</sup>, Chenlei Leng<sup>2</sup>, Ting Yan<sup>3</sup>, Qiwei Yao<sup>4</sup> and Xinyang Yu<sup>1</sup>

<sup>1</sup>The Hong Kong Polytechnic University

<sup>2</sup>University of Warwick

<sup>3</sup>Central China Normal University

<sup>4</sup>London School of Economics  
by.jiang@polyu.edu.hk

Analysis of networks that evolve dynamically requires the joint modelling of individual snapshots and time dynamics. This paper

## Abstracts

---

proposes a new flexible two-way heterogeneity model towards this goal. The new model equips each node of the network with two heterogeneity parameters, one to characterize the propensity to form ties with other nodes statically and the other to differentiate the tendency to retain existing ties over time. With  $n$  observed networks each having  $p$  nodes, we develop a new asymptotic theory for the maximum likelihood estimation of  $2p$  parameters when  $np \rightarrow \infty$  in which  $n \geq 2$  can be finite. We overcome the global non-convexity of the negative log-likelihood function by the virtue of its local convexity, and propose a novel method of moment estimator as the initial value for a simple algorithm that leads to the consistent maximum likelihood estimator (MLE). To establish the upper bounds for the estimation error of the MLE, we derive a new uniform deviation bound, which is of independent interest. The theory of the model and its usefulness are further supported by extensive simulation and a data analysis examining social interactions of ants.

### Estimating Conditional Covariance Matrices Dependent on Exogenous Variables

♦Hui Jiang and Jinze Ji

Huazhong University of Science and Technology  
jianghui@hust.edu.cn

Modelling and estimation of conditional covariance matrices play an important role in various fields. Numerous previous studies addressing covariance matrix estimation usually assume that the covariance matrix is constant or time-varying, but in reality, the relations between variables may be affected by some other factors. Therefore, we focus on the modelling and estimation of non-sparse conditional covariance matrix dependent on exogenous variables by introducing a semiparametric factor model structure.

### Session 23CHI139: Statistical Modeling for Neuroimaging Data

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

♦Ben Wu<sup>1</sup>, Keru Wu<sup>2</sup> and Jian Kang<sup>3</sup>

<sup>1</sup>Renmin University of China

<sup>2</sup>Duke University

<sup>3</sup>University of Michigan  
wuben@ruc.edu.cn

Deep neural networks (DNN) have been adopted in the scalar-on-image regression which predicts the outcome variable using image predictors. However, training DNN often requires a large sample size to achieve a good prediction accuracy and the model fitting results can be difficult to interpret. In this work, we construct a novel single-layer Bayesian neural network (BNN) with spatially varying weights for the scalar-on-image regression. Our goal is to select interpretable image regions and to achieve high prediction accuracy with limited training samples. We assign the soft-thresholded Gaussian process (STGP) prior to the spatially varying weights and develop an efficient posterior computation algorithm based on stochastic gradient Langevin dynamics (SGLD). The BNN-STGP provides large prior support for sparse, piecewise-smooth, and continuous spatially varying weight functions, enabling efficient posterior inference on image region selection and automatically determining the network structures. We establish the posterior consistency of model parameters and selection consistency of image regions when the number of voxels/pixels grows much faster than the sample size. We compared our methods with state-of-the-art deep learning methods

via analyses of multiple real data sets including the task fMRI data in the Adolescent Brain Cognitive Development (ABCD) study.

#### Minimizing Estimated Risks on Unlabeled Data for Semi-Supervised Medical Image Segmentation

♦Fuping Wu<sup>1</sup> and Xiahai Zhuang<sup>2</sup>

<sup>1</sup>University of Oxford

<sup>2</sup>Fudan University  
fuping\_wu@outlook.com

Supervised segmentation can be costly, particularly in applications of biomedical image analysis where large scale manual annotations from experts are generally too expensive to be available. Semi-supervised segmentation, able to learn from both the labeled and unlabeled images, could be an efficient and effective alternative for such scenarios. In this work, we propose a new formulation based on risk minimization, which makes full use of the unlabeled images. Different from most of the existing approaches which solely explicitly guarantee the minimization of prediction risks from the labeled training images, the new formulation also considers the risks on unlabeled images. Particularly, this is achieved via an unbiased estimator, based on which we develop a general framework for semi-supervised image segmentation. We validate this framework on three medical image segmentation tasks, namely cardiac segmentation on ACDC2017, optic cup and disc segmentation on REFUGE dataset and 3D whole heart segmentation on MM-WHS dataset. Results show that the proposed estimator is effective, and the segmentation method achieves superior performance and demonstrates great potential compared to the other state-of-the-art approaches.

#### Ensembled Seizure Detection Based on Small Training Samples

♦Peifeng Tong<sup>1</sup>, Haoxiang Zhan<sup>2</sup> and Songxi Chen<sup>3</sup>

<sup>1</sup>Guanghua School of Management, Peking University, Beijing 100871, China

<sup>2</sup>School of Mathematical Science, Peking University, Beijing 100871, China

<sup>3</sup>School of Mathematical Science and Guanghua School of Management, Peking University, Beijing 100871, China  
tongpf@pku.edu.cn

This paper proposes an interpretable ensembled seizure detection procedure based on the electroencephalographic (EEG) data, which integrates data driven features and expert knowledge while being robust against artifacts interference. The procedure is built on the spatially constrained independent component analysis supplemented by a knowledge based sparse seizure waveform representation to extract seizure intensity and waveform features, followed by a multiple change point detection algorithm to overcome the non-stationarity of the EEG signals and to perform temporal feature aggregation. The selected features are fed into a random forest classifier for ensembled seizure detection. Compared with the existing methods, the proposed procedure has the ability to identify seizure onset periods using only 5% of training samples (few-shot learning). Empirical performance on 24 patients in the CHB-MIT dataset showed superior performance of the proposed procedure.

#### A Semiparametric Gaussian Mixture Model for Chest Ct-Based 3d Blood Vessel Reconstruction

Qianhan Zeng

PEKING UNIVERSITY  
helenology@stu.pku.edu.cn

Computed tomography (CT) has been a powerful diagnostic tool since its emergence in the 1970s. Its highly detailed and high-resolution results contribute significantly to medical screening, especially for early stage lung cancer detection. Based on CT data,

the three-dimensional (3D) structures of human internal organs and tissues (e.g., blood vessels) can be reconstructed using professional software. 3D reconstruction is extremely beneficial for surgical operations and can serve as a vivid medical teaching example. However, traditional 3D reconstruction relies on manual operation by experienced surgeons, which is time-consuming, subjective, and requires substantial experience. To address this problem, we develop a novel semiparametric Gaussian mixture model for 3D blood vessel reconstruction. We theoretically extend the classical Gaussian mixture model by allowing both the component-wise mean and variance to nonparametrically vary according to voxel positions. A kernel-based expectation-maximization algorithm is developed to estimate the model, and a supporting asymptotic theory is established. A novel regression method is proposed for bandwidth selection, which is then compared with the traditional cross-validation-based method. The regression method outperforms the cross-validation method in both computational and statistical efficiency. In application, the 3D structures of blood vessels are successfully reconstructed in a fully automatic manner.

## Session 23CHI140: Network Data and Large-Scale Computing

### Subsampling-Based Modified Bayesian Information Criterion for Large-Scale Stochastic Block Models

*Jiayi Deng<sup>1</sup>, ♦Danyang Huang<sup>1</sup>, Xiangyu Chang<sup>2</sup> and Bo Zhang<sup>1</sup>*

<sup>1</sup>Renmin University of China

<sup>2</sup>Xi'an Jiaotong University  
dyhuang89@126.com

Identifying the number of communities is a fundamental problem in community detection, which has received increasing attention recently. However, rapid advances in technology have led to the emergence of large-scale networks in various disciplines, thereby making existing methods computationally infeasible. To address this challenge, we propose a novel subsampling-based modified Bayesian information criterion (SM-BIC) for identifying the number of communities in a network generated via the stochastic block model and degree-corrected stochastic block model. We first propose a node-pair subsampling method to extract an informative subnetwork from the entire network, and then we derive a purely data-driven criterion to identify the number of communities for the subnetwork. In this way, the SM-BIC can identify the number of communities based on the subsampled network instead of the entire dataset. This leads to important computational advantages over existing methods. We theoretically investigate the computational complexity and identification consistency of the SM-BIC. Furthermore, the advantages of the SM-BIC are demonstrated by extensive numerical studies.

### Statistical Analysis of Fixed Mini-Batch Gradient Descent Estimator

**♦Haobo Qi<sup>1</sup>, Feifei Wang<sup>2</sup> and Hansheng Wang<sup>1</sup>**

<sup>1</sup>Peking University

<sup>2</sup>Renmin University of China  
qihaoobo\_gsm@pku.edu.cn

We study here a fixed mini-batch gradient decent (FMGD) algorithm to solve optimization problems with massive datasets. In FMGD, the whole sample is split into multiple non-overlapping partitions. Once the partitions are formed, they are then fixed throughout the rest of the algorithm. For convenience, we refer to the fixed partitions as fixed mini-batches. Then for each computation iteration, the gradients are sequentially calculated on each fixed mini-

batch. Because the size of fixed mini-batches is typically much smaller than the whole sample size, it can be easily computed. It makes FMGD computationally efficient and practically more feasible. To demonstrate the theoretical properties of FMGD, we start with a linear regression model with a constant learning rate. We study its numerical convergence and statistical efficiency properties. We find that sufficiently small learning rates are necessarily required for both numerical convergence and statistical efficiency. Nevertheless, an extremely small learning rate might lead to painfully slow numerical convergence. To solve the problem, a diminishing learning rate scheduling strategy can be used. This leads to the FMGD estimator with faster numerical convergence and better statistical efficiency. Finally, the FMGD algorithms with random shuffling and a general loss function are also studied.

### Quasi-Score Matching Estimation for Spatial Autoregressive Model with Random Weights Matrix and Regressors

**♦Xuan Liang and Tao Zou**

The Australian National University  
xuan.liang@anu.edu.au

Due to the rapid development of technology to collect data, it becomes more often to apply the spatial autoregressive (SAR) model for a large size of dataset in real applications. However, the commonly used quasi-maximum likelihood estimation (QMLE) for the SAR model is not computationally scalable as the data size is large. In addition, when establishing the asymptotic properties of the parameter estimators of the SAR model, both weights matrix and regressors are assumed to be nonstochastic in classical spatial econometrics, which is perhaps not realistic in real applications. Motivated by the machine learning literature, this paper proposes quasi-score matching estimation for the SAR model. This new estimation approach is still likelihood-based, but significantly reduces the computational complexity of the QMLE. The asymptotic properties of parameter estimators under the random weights matrix and regressors are established, which provides a new theoretical framework for the asymptotic inference of the SAR-type models. The usefulness of the quasi-score matching estimation and its asymptotic inference are illustrated via extensive simulation studies and a case study of an anti-conflict social network experiment for middle school students.

### Identifying Temporal Pathways using Biomarkers in the Presence of Latent Non-Gaussian Components

**♦Shanghong Xie<sup>1</sup>, Donglin Zeng<sup>2</sup> and Yuanjia Wang<sup>3</sup>**

<sup>1</sup>Southwestern University of Finance and Economics

<sup>2</sup>University of North Carolina at Chapel Hill

<sup>3</sup>Columbia University  
shanghongxie@gmail.com

Time series data collected from a network of random variables are useful for identifying temporal pathways among the network nodes. Observed measurements may contain multiple sources of signals and noises, including Gaussian signals of interest and non-Gaussian noises including artifacts, structured noise, and other unobserved factors (e.g., genetic risk factors, disease susceptibility). Existing methods including vector autoregression (VAR) and dynamic causal modeling do not account for unobserved non-Gaussian components. Furthermore, existing methods cannot effectively distinguish contemporaneous relationships from temporal relations. In this work, we propose a novel method to identify latent temporal pathways using time series biomarker data collected from multiple subjects. The model adjusts for the non-Gaussian components and separates the temporal network from the contemporaneous network. Specif-

## Abstracts

ically, an independent component analysis (ICA) is used to extract the unobserved non-Gaussian components, and residuals are used to estimate the contemporaneous and temporal networks among the node variables based on method of moments. The algorithm is fast and can easily scale up. We derive the identifiability and the asymptotic properties of the temporal and contemporaneous networks. We demonstrate superior performance of our method by extensive simulations and an application to a study of attention-deficit/hyperactivity disorder (ADHD).

### Session 23CHI170: Checking Structural Change of Complex Data

#### Score Function-Based Tests for Ultrahigh-Dimensional Linear Models

♦ Weichao Yang, Xu Guo and Lixing Zhu

Beijing Normal University

20203101011@mail.bnu.edu.cn

To sufficiently exploit the model structure under the null hypothesis such that the conditions on the whole model can be mild, this paper investigates score function-based tests to check the significance of an ultrahigh-dimensional sub-vector of the model coefficients when the nuisance parameter vector is also ultrahigh-dimensional in linear models. We first reanalyze and extend a recently proposed score function-based test to derive, under weaker conditions, its limiting distributions under the null and local alternative hypotheses. As it may fail to work when the correlation between testing covariates and nuisance covariates is high, we propose an orthogonalized score function-based test with two merits: debiasing to make the non-degenerate error term degenerate and reducing the asymptotic variance to enhance the power performance. Simulations evaluate the finite-sample performances of the proposed tests, and a real data analysis illustrates its application.

#### Quantile Regression for Complex Longitudinal Data

♦ Xuerui Li, Yanyan Liu and Yuanshan Wu  
xqli@bnu.edu.cn

We develop a weighted quantile regression model for non-synchronous data. The estimate at the specific univariate quantile is obtained by minimizing the weighted objective loss function. We establish the asymptotic theory of the proposed estimators. The complicated structure results in the failure of using the classical cross-validation method to select the bandwidth of the kernel. To address this problem, we propose a data-driven bandwidth selection procedure. Using bootstrap to build a confidence interval.

#### Weighted Residual Empirical Processes, Martingale Transformations, and Model Checking for Regressions

♦ Falong Tan<sup>1</sup>, Xu Guo<sup>2</sup> and Lixing Zhu<sup>3</sup>

<sup>1</sup>Hunan University

<sup>2</sup>Beijing normal university

<sup>3</sup>Beijing normal university at Zhuhai  
falongtan@hnu.edu.cn

This paper propose a new methodology for testing the parametric forms of the mean and variance functions based on weighted residual empirical processes and their martingale transformations in regression models. The dimensions of the parameter vectors can be divergent as the sample size goes to infinity. We then study the convergence of weighted residual empirical processes and their martingale transformation under the null and alternative hypotheses in the diverging dimension setting. The proposed tests based on weighted residual empirical processes can detect local alternatives

distinct from the null at the fastest possible rate of order  $n^{1/2}$  but are not asymptotically distribution-free. While the tests based on martingale transformed weighted residual empirical processes can be asymptotically distribution-free, yet, unexpectedly, can only detect the local alternatives converging to the null at a much slower rate of order  $n^{1/4}$ , which is somewhat different from existing asymptotically distribution-free tests based on martingale transformations. As the tests based on the residual empirical process are not distribution-free, we propose a smooth residual bootstrap and verify the validity of its approximation in diverging dimension settings. Simulation studies and a real data example are conducted to illustrate the effectiveness of our tests.

#### Multiple Change Point Detection in Tensors

♦ Jiaqi Huang<sup>1</sup>, Junhui Wang<sup>2</sup>, Lixing Zhu<sup>3</sup> and Xuehu Zhu<sup>4</sup>

<sup>1</sup>Beijing Normal Univeisity

<sup>2</sup>Chinese University of Hong Kong

<sup>3</sup>Beijing Normal University

<sup>4</sup>Xi'an Jiaotong University

jhuang@mail.bnu.edu.cn

This paper proposes a criterion for detecting change structures in tensor data. To accommodate tensor structure with structural mode that is not suitable to be equally treated and summarized in a distance to measure the difference between any two adjacent tensors, we define a mode-based signal-screening Frobenius distance for the moving sums of slices of tensor data to handle both dense and sparse model structures of the tensors. As a general distance, it can also deal with the case without structural mode. Based on the distance, we then construct signal statistics using the ratios with adaptive-to-change ridge functions. The number of changes and their locations can then be consistently estimated in certain senses, and the confidence intervals of the locations of change points are constructed. The results hold when the size of the tensor and the number of change points diverge at certain rates, respectively. Numerical studies are conducted to examine the finite sample performances of the proposed method. We also analyze two real data examples for illustration.

### Session 23CHI25: Current Topics in Biostatistics II

#### Inference with Non-Differentiable Surrogate Loss in a General High-Dimensional Classification Framework

♦ Muxuan Liang<sup>1</sup>, Yang Ning<sup>2</sup>, Maureen Smith<sup>3</sup> and Ying-Qi Zhao<sup>4</sup>

<sup>1</sup>University of Florida

<sup>2</sup>Cornell University

<sup>3</sup>University of Wisconsin-Madison

<sup>4</sup>Fred Hutchinson Cancer Center  
muxuan.liang@ufl.edu

Penalized empirical risk minimization with a surrogate loss function is often used to derive a high-dimensional linear decision rule in classification problems. Although much of the literature focus on the generalization error, there is a lack of valid inference procedures to identify the driving factors of the estimated decision rule, especially when the surrogate loss is non-differentiable. In this work, we propose a kernel-smoothed decorrelated score to construct hypothesis testing and interval estimations for the linear decision rule estimated using a piece-wise linear surrogate loss, which has a discontinuous gradient and non-regular Hessian. Specifically, we adopt kernel approximations to smooth the discontinuous gradient near discontinuity points and approximate the non-regular Hessian of the

surrogate loss. In applications where additional nuisance parameters are involved, we propose a novel cross-fitted version to accommodate flexible nuisance estimates and kernel approximations. We establish the limiting distribution of the kernel-smoothed decorrelated score and its cross-fitted version in a high-dimensional setup. Simulation and real data analysis are conducted to demonstrate the validity and the superiority of the proposed method.

### **Bayesian Regression for Correlated Compositional Outcomes**

*Nanhua Zhang*

University of Cincinnati  
nanhua.zhang@cchmc.org

It is common to observe compositional data in many fields and we consider the compositional data as the outcome in a regression setting. We develop a Dirichlet regression model for the compositional outcome when the outcomes are correlated because of nesting within the same subject or a group. We motivate and apply the method proposed to data from a study of the effect of sleep restriction on physical activity outcomes, and we illustrate properties of the methods by simulation.

### **Standardization of Continuous and Categorical Covariates in Sparse Penalized Regressions**

*Xiang Li<sup>1</sup>, ♦Qing Pan<sup>2</sup> and Yong Ma<sup>3</sup>*

<sup>1</sup>Capital One

<sup>2</sup>George Washington University

<sup>3</sup>US FDA

qpan@gwu.edu

In sparse penalized regressions, candidate covariates of different units need to be standardized beforehand so that the coefficient sizes are directly comparable and reflect their relative impacts, which leads to fairer variable selection. However, when covariates of mixed data types (e.g. continuous, binary or categorical) exist in the same dataset, the commonly used standardization methods may lead to different selection probabilities even when the covariates have the same impact on or level of association with the outcome. In the paper, we propose a novel standardization method that targets at generating comparable selection probabilities in sparse penalized regressions for continuous, binary or categorical covariates with the same impact. We illustrate the advantages of the proposed method in simulation studies, and apply it to the National Ambulatory Medical Care Survey data to select factors related to the opioid prescription in US.

### **Session 23CHI29: Advanced Statistical Considerations on Contemporary Clinical Trials**

#### **Advancements and Challenges in Clinical Research Design for Rare Diseases**

*Hou Yan*

houyan@bjmu.edu.cn

TBC

#### **Mendelian Randomization for Drug Target Discovery**

*Xiao Xiang*

Fosun Pharma

xiangxiao@fosunpharma.com

The current drug discovery paradigm relies heavily on translational studies based on observational clinical study data. Although recent multi-omics data and the various deep learning methods have been adopted, the causality between the observed multi-omics features and risk of disease onset/progression have not caught adequate attention. Mendelian randomization using germline single

nucleotide polymorphism as instrumental variables has been extensively used in genetic epidemiology to infer the causal relationship between exposure and complex traits. The application of mendelian randomization may help pharmaceutical companies to repurposing marketed products to new indications. In addition, by applying mendelian randomization to individual or summary level data from large scale proteomics studies, novel drug targets may be discovered. In this talk, I will first summarize the advantages and pitfalls of current statistical methods in mendelian randomization. I will then introduce the proposed analysis workflows in drug repurposing and target discovery. I will also introduce a case study in repurposing a respiratory neutrophil targeted therapy to autoimmune disease using Mendelian randomization in transcriptome and proteomics of serum proteins.

### **Phase II Trial Design Based on Success Probabilities for Phase III in Oncology: a Case Study**

*Xun Liao*

默克雪兰诺（北京）医药研发有限公司  
XUN.A.LIAO@MERCKGROUP.COM

In recent years, high failure rates in phase III trials were observed. One of the main reasons is overoptimistic assumptions for the planning of phase III resulting from limited phase II information and/or unawareness of realistic success probabilities. Heiko Gotte, et al in 2015 provided an approach for planning a phase II trial in a time-to-event setting that considers the whole phase II/III clinical development program. In this approach, the simulations results show that unconditional probabilities of go/no-go decision as well as the unconditional success probabilities for phase III are influenced by the number of events observed in phase II. However, choosing more than 150 events in phase II seems not necessary as the impact on these probabilities then becomes quite small. Here we would like to illustrate the application of this approach on the study design of a phase II trial through a case study in a time-to-event setting.

### **Session 23CHI61: Recent Methodological Advances in Survey Statistics**

#### **A Multivariate Bayesian Hierarchical Model for Small Area Estimation of Criminal Victimization Rates in Domains Defined by Age and Gender**

*Emily Berg*

Iowa State University

emilyb@iastate.edu

The National Crime Victimization Survey (NCVS) gathers information on criminal victimizations for individuals in a representative sample of United States households. The NCVS provides authoritative data on the rates of many types of violent crimes, including simple assault, robbery, and aggravated assault. The NCVS is designed to permit direct estimation for 22 large states and for the nation. Small sample sizes preclude production of direct estimates for more detailed domains. Past work on small area estimation with the NCVS data has focused heavily on geographic subdivisions of the United States. Estimates for demographic subdivisions are also of interest. We consider the problem of producing estimates for small domains defined by the intersection of age categories and genders. We construct estimates for four types of violent crimes in each of two time periods. We accomplish this through the use of a multivariate Bayesian hierarchical model. We compare a model with a log transformation to a model fit to the data in the original scale.

## Abstracts

---

We compare small area predictors based on a selected model to the direct estimators.

### A-Optimal Split Questionnaire Designs

Zhengyuan Zhu

Iowa State University

zhuz@iastate.edu

One of the main challenges in survey data collection is the low response rate, particularly for long surveys. Split Questionnaire Designs (SQD) address this issue by randomly assigning a fraction of the full questionnaires to respondents, thus reducing response burden and potentially increasing response rates and improving survey quality. Traditional SQD methods divide questions into a small number of panels based on topics, assigning each respondent a subset of the panels to complete. In this talk, we present a general probabilistic framework for determining the optimal SQD (OSQD) based on a given optimality criterion, which allows for more flexible question selection to minimize information loss. Our methods utilize prior survey data to calculate the Fisher information matrix and apply the A-optimality criterion to identify the OSQD for the current survey study. We conduct simulation studies comparing OSQDs to baselines and apply our method to the 2016 Pet Demographic Survey (PDS) data. In both simulation studies and real data application, local and global OSQDs outperform other baseline methods.

### Augmented Two-Step Estimating Equations with Nuisance Functionals and Complex Survey Data

♦Puying Zhao<sup>1</sup> and Changbao Wu<sup>2</sup>

<sup>1</sup>Yunnan University

<sup>2</sup>University of Waterloo

pyzhao@live.cn

Statistical inference in the presence of nuisance functionals with complex survey data is an important topic in social and economic studies. The Gini index, Lorenz curves and quantile shares are among the commonly encountered examples. The nuisance functionals are usually handled by a plug-in nonparametric estimator and the main inferential procedure can be carried out through a two-step generalized empirical likelihood method. Unfortunately, the resulting inference is not efficient and the nonparametric version of the Wilks' theorem breaks down even under simple random sampling. We propose an augmented estimating equations method with nuisance functionals and complex surveys. The second-step augmented estimating functions obey the Neyman orthogonality condition and automatically handle the impact of the first-step plug-in estimator, and the resulting estimator of the main parameters of interest is invariant to the first step method. More importantly, the generalized empirical likelihood based Wilks' theorem holds for the main parameters of interest under the design-based framework for commonly used survey designs, and the maximum generalized empirical likelihood estimators achieve the semiparametric efficiency bound. Performances of the proposed methods are demonstrated through simulation studies and an application using the dataset from the New York City Social Indicators Survey.

### Session 23CHI94: Modern Statistical Process Control and Change-Point Problems II

#### Phase Ii Control Chart Based on Likelihood Ratio Test for Monitoring Time Between Events of Gumbel's Bivariate Exponential Distribution

♦Jiujun Zhang and Peile Chen

zjjly790816@163.com

Schemes for monitoring two or more time between events data are becoming increasingly significant in statistical process monitoring. In practice, parameters must be estimated prior to the start of process monitoring and control. In this paper, we first propose a multivariate time between events control chart based on the one-sample log-likelihood ratio test for monitoring the Gumbel's bivariate exponential (GBE) distribution, and then we analyze the effect of parameter estimation on this chart. To avoid the effect of parameter estimation, we then propose a method based on the two-sample log-likelihood ratio test for monitoring the shifts occurring in either the scale parameter or the dependence parameter or both for the GBE distribution process. The proposed method was designed with two-sample log-likelihood ratio tests for univariate marginal distribution functions and Gumbel copula function. A Max-type function-based approach was implemented for joint monitoring of margin distributions and copula-related structures. In contrast to the traditional schemes used to monitor bivariate processes, the strength of our proposed technique lies in the ability to identify exactly which component of the scale parameter or dependence parameter is responsible for the source of the signal. The performance of the proposed chart was analyzed via Monte Carlo simulations.

#### Low-Rank Matrix Estimation in the Presence of Change-Points

Lei Shi<sup>1</sup>, ♦Guanghui Wang<sup>2</sup> and Changliang Zou<sup>3</sup>

<sup>1</sup>University of California, Berkeley

<sup>2</sup>East China Normal University

<sup>3</sup>Nankai University

ghwang.nk@gmail.com

We consider a general trace regression model with multiple structural changes, and propose a universal approach for simultaneous exact or near low-rank matrix recovery and changepoint detection. It incorporates nuclear norm penalized least-squares minimization into a grid search scheme that determines the potential structural break. Under a set of general conditions, we establish the non-asymptotic error bounds with a nearly-oracle rate for the matrix estimators as well as the super-consistency rate for the change-point localization. We use concrete random design instances to justify the appropriateness of the proposed conditions. Numerical results demonstrate the validity and effectiveness of the proposed scheme.

#### Identification of Outlying Observations for Large-Dimensional Data

Tao Wang<sup>1</sup>, Xiaona Yang<sup>2</sup>, Yunfei Guo<sup>3</sup> and ♦Zhonghua Li<sup>3</sup>

<sup>1</sup>Huaiyin Normal University

<sup>2</sup>Heilongjiang University

<sup>3</sup>Nankai University

zli@nankai.edu.cn

In this talk, we will propose a two-stage procedure for identifying outlying observations in large-dimensional data set. In the first stage, an outlier identification measure is defined by using a max-normal statistic and a clean subset that contains non-outliers is obtained. The identification of outliers can be deemed as a multiple hypothesis testing problem, then, in the second stage, we explore the asymptotic distribution of the proposed measure, and obtain the threshold of the outlying observations. Furthermore, in order to improve the identification power and better control the misjudgment rate, a one step refined algorithm is proposed. Simulation results and real data analysis examples show that, compared with other methods, the proposed procedure has great advantages in identifying outliers in various data situations.

## Dynamic Modeling and Online Monitoring of Tensor Data Streams with Application to Passenger Flow Surveillance

*Wendong Li*

Shanghai University of Finance and Economics  
wendongli01@gmail.com

Dynamic tensor data streams are becoming prevalent in various application domains such as traffic networks, smart manufacturing, etc. It is crucial to monitor such tensor data streams to detect abnormal activities and system failures promptly. Existing tensor monitoring methods either rely heavily on the assumption that the tensor coefficients exhibit a low-rank structure or are inapplicable to general-order tensors. In this article, prompted by the surveillance of subway passenger flow, we propose a unified framework for the dynamic modeling and online monitoring of tensor data streams. Based on the tensor normal distribution, we first derive a tensor model selection procedure, through which the selected tensor structure strikes a balance between model complexity and estimation accuracy. Then we propose an online estimation procedure to estimate the parameters of the tensor process dynamically, based on which sequential change-detection procedures are proposed using the generalized likelihood ratio test. By benefiting from modeling the complex correlation structure of different tensor modes dynamically, our procedures can improve the sensitivity with which various types of changes are detected by comparison with existing methods. The efficacy of our approach is illustrated by extensive simulations and an analysis of real passenger flow surveillance data.

## Session 23CHI132: Advanced Methods in Adaptive Randomization Designs

### Discussion: New Developments in Adaptive Randomization Designs

*Li-Xin Zhang*

Zhejiang University  
stazlx@zju.edu.cn

Adaptive randomization designs have been proven to be advantageous in providing more efficient clinical trials, increasing the likelihood of receiving better treatment, and balancing covariates. Recent studies are focusing on complex circumstances such as missing responses, continuous covariates, heteroscedastic covariates, unobserved covariates, etc. It is a challenge to develop new adaptive designs for randomizing patients in an efficient way so that the statistical inference is valid, efficient, and robust.

### Model-Based Adaptive Randomization Procedures for Heteroscedasticity of Treatment Responses

*Zhongqiang Liu*

Henan Polytechnic University  
zhongqiang@hpu.edu.cn

In clinical trials, the responses of patients usually depend on the assigned treatment as well as some important covariates, which may cause heteroscedasticity in treatment responses. As clinical trials are generally designed to demonstrate efficacy for the overall population, they are usually not adequately powered for detecting interactions. To improve the power of interaction tests, this article develops two model-based adaptive randomization procedures for heteroscedasticity of treatment responses, and derives their limiting allocation proportions, which are generalizations of the Neyman allocation. Issues of hypothesis testing and sample size estimation are also addressed. Simulation studies show that compared with complete randomization, the two model-based randomization pro-

cedures have greater power to detect differences in systematic effects, main treatment effects and treatment–covariate interactions. In addition, the validity of limiting allocation proportion is also verified through simulations.

### A New and Unified Family of Covariate Adaptive Randomization Procedures and their Properties

♦Wei Ma<sup>1</sup>, Ping Li<sup>2</sup>, Li-Xin Zhang<sup>3</sup> and Feifang Hu<sup>4</sup>

<sup>1</sup>Renmin University of China

<sup>2</sup>LinkedIn Corporation

<sup>3</sup>Zhejiang University

<sup>4</sup>George Washington University  
mawei@ruc.edu.cn

In clinical trials and other comparative studies, covariate balance is crucial for credible and efficient assessment of treatment effects. Covariate adaptive randomization (CAR) procedures are extensively used to reduce the likelihood of covariate imbalances occurring. In the literature, most studies have focused on balancing of discrete covariates. Applications of CAR with continuous covariates remain rare, especially when the interest goes beyond balancing only the first moment. In this talk, we propose a family of CAR procedures that can balance general covariate features, such as quadratic and interaction terms. Our framework not only unifies many existing methods, but also introduces a much broader class of new and useful CAR procedures. We show that the proposed procedures have superior balancing properties; in particular, the convergence rate of imbalance vectors is  $O_P(n^\epsilon)$  for any  $\epsilon > 0$  if all of the moments are finite for the covariate features, relative to  $O_P(\sqrt{n})$  under complete randomization, where  $n$  is the sample size. Both the resulting convergence rate and its proof are novel. These favorable balancing properties lead to increased precision of treatment effect estimation in the presence of nonlinear covariate effects. The framework is applied to balance covariate means and covariance matrices simultaneously.

### Balancing Unobserved Covariates with Covariate-Adaptive Randomized Experiments

♦Yang Liu<sup>1</sup> and Feifang Hu<sup>2</sup>

<sup>1</sup>Renmin University of China

<sup>2</sup>George Washington University  
yangliu2022@ruc.edu.cn

Establishing balance for important covariates is often critical in causal inference and clinical trials. Covariate-adaptive randomization (CAR) and stratified permuted block (STR-PB) are commonly implemented to achieve this goal. While the balance properties of these methods have been extensively studied for observed covariates, their properties for balancing unobserved covariates have less been understood from the theoretical aspect, and have been subjected to criticism in the literature. In this presentation, we will introduce a framework for assessing the theoretical properties of unobserved covariate imbalance. This versatile framework allows for the analysis and the comparison for the balance properties of complete randomization (CR), STR-PB, and other CAR procedures in relation to the unobserved covariates. Our findings highlight the advantages of utilizing CAR or STR-PB (especially when the number of strata is relatively small compared to the sample size) for balancing unobserved covariates. Numerical studies are also presented to demonstrate finite sample properties of our theoretical results. These findings not only serve as a foundation for the future research on the impact of unobserved covariates for covariate-adaptive randomized trials but also open up possibilities for various other applications.

## Abstracts

---

### Session 23CHI14: New Developments in Nonparametric and Semiparametric Methods for Complex Data

#### Latent Single-Index Models with Factor Structure for Multivariate Ordinal Data

Zhiyong Chen

School of Mathematics and Statistics, Fujian Normal University  
zychen1024@163.com

We propose a general latent semi-parametric model for multivariate ordinal data in which the single-index models with factor structure are used to assess the effects of the latent covariates on the latent responses and to explore the covariance structure of the latent responses. Since the indices may vary from minus infinity to plus infinity, the traditional spline can not be applied directly to approximate the unknown link function. We employ a modified version to tackle this pervasive problem by first transforming the indices into the unit interval via a continuously cumulative distribution function and then constructing a novel combination of the spline bases on the unit interval. A fully Bayesian modelling is performed by facilitating efficient Markov chain Monte Carlo approach with free-knot splines to analyze the proposed model. To accelerate the convergence, we make use of the partial-collapse and parameter expansion and reparameterization techniques and design a generalized Metropolis step in our algorithm. The performance of the proposed model and estimation method are checked through a simulation study and applied to a real dataset.

#### Electricity Consumption Forecasting by a New Neural Network Model: Panel Semiparametric Quantile Regression Neural Network (Psqrnn)

Xingcai Zhou, ♦Jiangyan Wang, Hongxia Wang and Jinguan Lin  
Nanjing Audit University  
wangjiangyan2007@126.com

Addressing the forecasting issues is one of the core objectives of developing and restructuring of electric power industry in China, however, no enough effort has been made to develop an accurate electricity consumption forecasting procedure. Motivated by which, a panel semiparametric quantile regression neural network (PSQRNN) is developed by combining an artificial neural network and semiparametric quantile regression for panel data. By combining the penalized quantile regression with least absolute shrinkage and selection operator (LASSO), ridge regression and backpropagation, PSQRNN keeps the flexibility of nonparametric models and the interpretability of parametric models simultaneously. The prediction accuracy of the proposed PSQRNN is evaluated based on China's electricity consumption data set, and the results indicate that PSQRNN performs better compared with three benchmark methods including BP neural network (BP), Support Vector Machine (SVM) and Quantile Regression Neural Network (QRNN).

#### Compositional Multivariate Regression for Microbiome Data Integration

Chenxiao Hu, Ying Dai, Thomas Sharpton and ♦Duo Jiang  
Oregon State University  
duo.jiang@oregonstate.edu

The emergence of microbiome multi-omics studies calls for effective statistical methods to infer the effects of microbiome measures on another omics data type. Penalized multivariate regression has proven a useful approach for integrating disparate types of omics data that both high-dimensional. However, existing methods fail to account for the compositionality of microbiome data. Specifically, using microbiome sequencing, the absolute abundances (AA) of microbes are unobservable and microbiome composition is only char-

acterized by the relative abundance (RA) of a microbe relative to other microbes. To resolve this challenge, we propose a novel multivariate regression framework that models the associations between microbiome data and a paired omic data construct. By capitalizing on the high-dimensionality of the data, our approach features the ability to assess the effects of microbiome absolute abundances by using relative abundance data only. Our model explicitly incorporates the unknown total microbial abundance into a reduced rank regression model with nuclear penalty. An ADMM-based algorithm is developed to estimate the microbiome-response association matrix. The advantages of our methods are shown in both simulation studies and an application to a zebrafish study.

#### Integrated Subgroup Identification from Multi-Source Data

Lihui Shao, ♦Jiaqi Wu, Weiping Zhang and Yu Chen

University of Science and Technology of China  
zwp@ustc.edu.cn

Subgroup identification is crucial in dealing with the heterogeneous population and has wide applications in various areas, such as clinical trials and market segmentation. With the prevalence of multi-source data, there is a practical need to identify subgroups based on multi-source data. This paper proposes a working-independence pseudo-loglikelihood and integrates the parameters of each source into a pairwise fusion penalty for simultaneous parameter estimation and subgroup identification. We derive an alternating direction method of multipliers (ADMM) algorithm for its implementation. Furthermore, the weak oracle properties of parameter estimation are established, illustrating the latent subgroups can be consistently identified. Finally, we conduct numerical simulations and the analysis of a randomized trial on reduced nicotine standards for cigarettes to evaluate the performance of the proposed method.

### Session 23CHI143: Recent Developments of Statistical Methods on Missing Data with Applications

#### A Self-Censoring Model for Multivariate Nonignorable Nonmonotone Missing Data

♦Yilin Li<sup>1</sup>, Wang Miao<sup>1</sup>, Ilya Shpitser<sup>2</sup> and Eric Tchetgen Tchetgen<sup>3</sup>

<sup>1</sup>Peking University

<sup>2</sup>Johns Hopkins University

<sup>3</sup>The Wharton School of the University of Pennsylvania  
yilinli@pku.edu.cn

We introduce an itemwise modeling approach called self-censoring" for multivariate nonignorable nonmonotone missing data, where the missingness process of each outcome can be affected by its own value and associated with missingness indicators of other outcomes, while conditionally independent of the other outcomes. The self-censoring model complements previous graphical approaches for the analysis of multivariate nonignorable missing data. It is identified under a completeness condition stating that any variability in one outcome can be captured by variability in the other outcomes among complete cases. For estimation, we propose a suite of semiparametric estimators including doubly robust estimators that deliver valid inferences under partial misspecification of the full-data distribution. We also provide a novel and flexible global sensitivity analysis procedure anchored at the self-censoring. We evaluate the performance of the proposed methods with simulations and apply them to analyze a study about the effect of highly active antiretroviral therapy on preterm delivery of HIV-positive mothers.

### Bayesian Diagnostics of Hidden Markov Structural Equation Models with Missing Data

♦*Jingheng Cai*<sup>1</sup>, *Ming Ouyang*<sup>2</sup>, *Kai Kang*<sup>1</sup> and *Xinyuan Song*<sup>3</sup>

<sup>1</sup>Sun Yat-sen University

<sup>2</sup>The Chinese University of Hong Kong

<sup>3</sup>The Chinese University of Hong Kong,  
caijheng@mail.sysu.edu.cn

In this study, we develop a Bayesian local influence procedure for HMMs with latent variables in the presence of missing data. The proposed model enables us to investigate the dynamic heterogeneity of multivariate longitudinal data, reveal how the interrelationships among latent variables change from one state to another, and simultaneously conduct statistical diagnosis for the given data, model assumptions, and prior inputs. We apply the proposed procedure to analyze a dataset collected by the UCLA center for advancing longitudinal drug abuse research. Several outliers or influential points that seriously influence estimation results are identified and removed. The proposed procedure also discovers the effects of treatment and individuals' psychological problems on cocaine use behavior and delineates their dynamic changes across the cocaine-addiction states.

### Bayesian Modeling and Inference for Item Response Model with Nonignorable Missing Data

♦*Zhihua Ma*<sup>1</sup>, *Jing Wu*<sup>2</sup> and *Ming-Hui Chen*<sup>3</sup>

<sup>1</sup>Shenzhen University

<sup>2</sup>The University of Rhode Island

<sup>3</sup>University of Connecticut  
mazh@szu.edu.cn

Not-reached (dropout) and omitted (intermittent missingness) items are often inevitable in timed tests where answers are not required. The missingness of the item response may be related to the subject's latent characteristics, the difficulty, or even the unobserved response of the item. To fully understand the underlying results of the testing, we must handle the missing data appropriately. In this article, a multilevel IRT model is built with the response time model as well as the missing data mechanisms to take account of the impact of response time and missingness. A new missing data mechanism model is proposed to jointly studies the not-reached and omitted behaviors. In order to assess the impact of the missing data mechanism model on the multilevel IRT model, a decomposition of the LPML criteria is used for model comparison. This proposed methodology is illustrated using extensive simulations and a real data from the Program for International Student Assessment (PISA) 2018 study.

### Session 23CHI157: Statistical Advances in Biomedical Data Analysis

#### A Hybrid Machine Learning and Regression Method for Cell Type Deconvolution of Spatial Barcoding-Based Transcriptomic Data

*Yunqing Liu*<sup>1</sup>, ♦*Ningshan Li*<sup>2</sup>, *Ji Qi*<sup>3</sup>, *Gang Xu*<sup>4</sup>, *Jiayi Zhao*<sup>3</sup>, *Nating Wang*<sup>3</sup>, *Aurélien Justet*<sup>5</sup>, *Taylor Adams*<sup>5</sup>, *Zuoheng Wang*<sup>3</sup> and *Xiting Yan*<sup>5</sup>

<sup>1</sup>Department of Biostatistics, Yale School of Public Health

<sup>2</sup>The Second Affiliated Hospital, School of Medicine, The Chinese University of Hong Kong, Shenzhen & Longgang District People's Hospital of Shenzhen

<sup>3</sup>Department of Biostatistics, Yale School of Public Health

<sup>4</sup>Department of Mathematical Sciences, University of Nevada

<sup>5</sup>Section of Pulmonary, Critical Care and Sleep Medicine, Yale School of Medicine  
hill103.2@gmail.com

Spatial barcoding-based transcriptomic (ST) data require cell type deconvolution for cellular-level downstream analysis. Here we present SDePER, a hybrid machine learning and regression method, to deconvolve ST data using reference single-cell RNA sequencing (scRNA-seq) data. SDePER, for the first time, removes the systematic difference between the ST and scRNA-seq data (platform effects) efficiently to ensure the linear relationship between ST data and cell type-specific expression profile. It also considers sparsity of cell types per capture spot and across-spots spatial correlation in cell type compositions. Based on the estimations, SDePER imputes for cell type compositions and gene expression at enhanced resolution. We assessed the performance of SDePER and six existing methods using simulations and four real datasets. All results showed that SDePER achieved significantly more accurate and robust results than the existing methods suggesting the importance of considering platform effects, sparsity and spatial correlation in cell type deconvolution.

#### Dimensionality Reduction with Network Regularization in Single-Cell Expression Analysis

*Huaying Fang*

Academy for Multidisciplinary Studies, Capital Normal University  
hyfang@cnu.edu.cn

Single-cell RNA sequencing (scRNA-seq) technology can measure gene expression abundance in single cells. However, there are two challenges for analyzing scRNA-seq data owing to technical limitations and expense budes. One challenge is the dropout problem that read counts of many genes are zeros in individual cells while the other challenge is that the number of genes is often much larger than the sample size. In this talk, we will introduce a novel method for simultaneously imputing zeros and reducing data dimensions for analyzing scRNA-seq data. The proposed method integrates the prior gene-gene interaction network with the observed scRNA-seq data. We will show that the proposed method outperforms existing methods for assisting clustering cells in most cases using both simulated data and gene expression profiles from scRNA-seq experiments.

#### Time-Varying Treatment Effects of Functional Data with Latent Confounders: Application to Sleep Heart Health Studie

♦*Jie Li*<sup>1</sup>, *Shujie Ma*<sup>2</sup> and *Yehua Li*<sup>2</sup>

<sup>1</sup>Renmin University of China

<sup>2</sup>University of California, Riverside  
lijie\_stat@ruc.edu.cn

Exploring the causal effect between variables is an important issue in lots of scientific research. Existing literature on causal inference mainly studies one-dimensional or multi-dimensional data, but functional data with repeated observations per individual frequently appears in a wide variety of applications. In functional data, treatment design may change across time, and its treatment effect is a time-varying function. Besides, most methods for treatment effect estimation based on observational data rely on the ignorability assumption that treatment assignment is independent of the potential outcomes given the observable covariates. This assumption can be violated when unobserved latent covariates are involved. We propose a novel method for unbiased treatment effect estimation with unobserved latent covariates for functional data. We propose to solve this challenging problem using a joint likelihood method with a Monte Carlo EM algorithm. Moreover, our proposed method is flexible to estimate both heterogeneous treatment effect of individuals and average treatment effect, providing a reliable inferential tool

## Abstracts

---

in making treatment decisions. It can also be applied to the irregular and sparse data. The method leads to meaningful discoveries when utilized to investigate the dynamic effect of sleep quality on heart rate variability.

### A Unified Quantile Framework Reveals Nonlinear Heterogeneous Transcriptome-Wide Associations

♦Tianying Wang<sup>1</sup>, Iuliana Ionita-Laza<sup>2</sup> and Ying Wei<sup>2</sup>

<sup>1</sup>Tsinghua University

<sup>2</sup>Columbia University

tianyingw0905@outlook.com

Transcriptome-wide association studies (TWAS) are powerful tools for identifying putative causal genes by integrating genome-wide association studies and gene expression data. Most existing methods are based on linear models and therefore may miss or underestimate nonlinear associations. In this article, we propose a robust, quantile-based, unified framework to investigate nonlinear transcriptome-wide associations in a quantile process manner. Through extensive simulations and the analysis of multiple psychiatric and neurodegenerative disorders, we showed that the proposed framework gains substantial power over conventional approaches and leads to insightful discoveries on nonlinear associations between gene expression levels and traits, thereby providing a complementary approach to existing literature. In doing so, we applied the proposed method for 797 continuous traits from the UK Biobank, and the results are available in a public repository.

### Session 23CHI158: Statistical Inferences in Computational Biology and Genetics

#### A Data-Adaptive Bayesian Regression Approach for Polygenic Risk Prediction

Lin Hou

Tsinghua University  
hou@tsinghua.edu.cn

Polygenic risk score (PRS) has been widely exploited for genetic risk prediction due to its accuracy and conceptual simplicity. We introduce a unified Bayesian regression framework, NeuPred, for PRS construction, which accommodates varying genetic architectures and improves overall prediction accuracy for complex diseases by allowing for a wide class of prior choices. To take full advantage of the framework, we propose a summary statistics-based cross-validation strategy to automatically select suitable chromosome-level priors, which demonstrates a striking variability of the prior preference of each chromosome, for the same complex disease, and further significantly improves the prediction accuracy. Simulation studies and real data applications with seven disease datasets from the Wellcome Trust Case Control Consortium cohort and eight groups of large-scale genome-wide association studies demonstrate that NeuPred achieves substantial and consistent improvements in terms of predictive  $r^2$  over existing methods. In addition, NeuPred has similar or advantageous computational efficiency compared with the state-of-the-art Bayesian methods.

#### Fastancom: a Fast Method for Analysis of Compositions of Microbiomes

Tao Wang

Shanghai Jiao Tong University  
neowangtao@sjtu.edu.cn

Analysis of compositions of microbiomes (ANCOM) compares the absolute abundances of microbes between two or more ecosystems

using relative abundances in specimens derived from these ecosystems. Despite its impressive performance, there are two drawbacks to ANCOM. First, with K microbes it requires fitting  $K(K-1)/2$  models for log-ratios of counts, and so can be computationally intensive. Second, it does not output P-values for microbes detected as differentially abundant. We propose a fast implementation of ANCOM, fastANCOM, that fits only K models for log-transformed counts. fastANCOM provides P-values to declare statistical significance and outputs log fold changes of abundance between groups. fastANCOM compares favorably with existing differential abundance testing methods.

#### Feature Screening for Clustering Analysis with Applications Single-Cell Rna Sequencing

Changhu Wang, Zihao Chen and ♦Ruixin Xi

Peking University

ruixinxi@math.pku.edu.cn

We consider feature screening for ultrahigh dimensional clustering analyses. Based on the observation that the marginal distribution of any given feature is a mixture of its conditional distributions in different clusters, we propose to screen clustering features by independently evaluating the homogeneity of each feature's mixture distribution. Important clustering-relevant features have heterogeneous components in their mixture distributions and unimportant features have homogeneous components. The well-known EM-test statistic is used to evaluate the homogeneity. Under general parametric settings, we establish the tail probability bounds of the EM-test statistic for the homogeneous and heterogeneous features, and further show that the proposed screening procedure can achieve the sure independent screening and even the consistency in selection properties. Limiting distribution of the EM-test statistic is also obtained for general parametric distributions. The proposed method is computationally efficient, can accurately screen for important clustering-relevant features and help to significantly improve clustering, as demonstrated in our extensive simulation and real data analyses. Applications in single-cell RNA sequencing are also discussed.

#### Robust and Powerful Gene-Environment Interaction Tests using Rare Genetic Variants in Case-Control Studies

♦Yanan Zhao and Hong Zhang

zhangh@ustc.edu.cn

Many association analysis methods have been developed to detect disease related rare genetic variants or gene-environment interactions. Most of them are based on a prospective likelihood, so they are robust but might not be powerful enough. On the other hand, retrospective likelihood based methods assuming gene-environment independence can effectively improve the power of association test, but they suffer from type-I error rate inflation if the independence assumption is violated. We aim to develop novel test methods to balance power and robustness by appropriately weighting retrospective likelihood based tests and prospective likelihood based tests. The desired finite sample performances of the proposed methods are demonstrated through simulation studies and the application to a real dataset.

### Session 23CHI18: Advanced Statistical Methods for Complex Data

#### Moment Deviation Subspaces of Dimension Reduction for High-Dimensional Data with Change Structure

♦Xuehu Zhu<sup>1</sup>, Luoyao Yu<sup>1</sup>, Jiaqi Huang<sup>2</sup>, Junmin Liu<sup>1</sup> and Lixing

Zhu<sup>2</sup>

<sup>1</sup>Xi'an Jiaotong University

<sup>2</sup>Beijing Normal University  
zhuxuehu@xjtu.edu.cn

This paper introduces the notion of moment deviation subspaces of dimension reduction for high-dimensional data with change structure. We propose a novel estimation method to identify subspaces by combining the Mahalanobis matrix and the pooled covariance matrix. The theoretical properties are investigated to show that the change point detection and clustering can be equivalently implemented in the dimension reduction subspaces, whether the data structure is dense or sparse, whenever the dimension divided by the sample size goes to zero. We propose an iterative algorithm based on dimension reduction subspaces that can be applied for data clustering of high-dimensional data. The numerical studies on synthetic and real data sets suggest that the dimension reduction versions of existing methods of change point detection and clustering methods significantly improve the performances of existing approaches in finite sample scenarios.

#### Health Utility Survival for Randomized Clinical Trials: a Composite Endpoint for Clinical Trial Designs

Yangqing Deng<sup>1</sup>, ♦Meiling Hao<sup>2</sup>, John R. De Almeida<sup>3</sup>, Xiaolu Wang<sup>2</sup> and Wei Xu<sup>4</sup>

<sup>1</sup>Princess Margaret Cancer Centre

<sup>2</sup>University of International Business and Economics

<sup>3</sup>University Health Network

<sup>4</sup>Princess Margaret Cancer Centre; University of Toronto  
meilinghao@uibe.edu.cn

Overall survival has been used as the primary endpoint for many randomized trials that aim to examine whether a new treatment is non-inferior to the standard treatment or placebo control. When a new treatment is indeed non-inferior in terms of survival, it may be important to assess other outcomes including health utility. However, analyzing health utility scores in a secondary analysis may have limited power since the primary objectives of the original study design may not include health utility. To comprehensively consider both survival and health utility, we developed a composite endpoint, HUS (Health Utility-adjusted Survival), which combines both survival and utility. With a detailed framework to conduct sample size calculation and power analysis, HUS has been shown to be able to increase statistical power and potentially reduce the required sample size compared to the standard overall survival endpoint. Nevertheless, the asymptotic properties of the test statistics of HUS endpoint have yet to be fully established. Besides that, the standard version of HUS cannot be applied to or have limited performance in certain scenarios, where extensions are needed. In this manuscript, we propose various methodological extensions of HUS and derive the asymptotic distributions of the test statistics. By comprehensive simulation

#### Statistical Inference in High-Dimensional Regression with Streaming Data

♦Ruijian Han<sup>1</sup>, Lan Luo<sup>2</sup>, Yuanyuan Lin<sup>3</sup> and Jian Huang<sup>1</sup>

<sup>1</sup>The Hong Kong Polytechnic University

<sup>2</sup>University of Iowa

<sup>3</sup>The Chinese University of Hong Kong  
ruijian.han@polyu.edu.hk

We propose a debiased stochastic gradient descent algorithm for online statistical inference with high-dimensional data. Our approach combines the debiasing technique developed in high-dimensional statistics with the stochastic gradient descent algorithm. It can be used for efficiently constructing confidence intervals in an online

fashion. Our proposed algorithm has several appealing aspects: first, as a one-pass algorithm, it reduces the time complexity; in addition, each update step requires only the current data together with the previous estimate, which reduces the space complexity. We establish the asymptotic normality of the proposed estimator under mild conditions on the sparsity level of the parameter and the data distribution. We conduct numerical experiments to demonstrate the proposed debiased stochastic gradient descent algorithm reaches nominal coverage probability.

#### Quantile Autoregressive Conditional Heteroscedasticity

♦Qianqian Zhu<sup>1</sup>, Songhua Tan<sup>1</sup>, Yao Zheng<sup>2</sup> and Guodong Li<sup>3</sup>

<sup>1</sup>Shanghai University of Finance and Economics

<sup>2</sup>University of Connecticut

<sup>3</sup>University of Hong Kong

zhu.qianqian@mail.shufe.edu.cn

This talk proposes a novel conditional heteroscedastic time series model by applying the framework of quantile regression processes to the ARCH() form of the GARCH model. This model can provide varying structures for conditional quantiles of the time series across different quantile levels, while including the commonly used GARCH model as a special case. The strict stationarity of the model is discussed. For robustness against heavy-tailed distributions, a self-weighted quantile regression (QR) estimator is proposed. While QR performs satisfactorily at intermediate quantile levels, its accuracy deteriorates at high quantile levels due to data scarcity. As a remedy, a self-weighted composite quantile regression (CQR) estimator is further introduced and, based on an approximate GARCH model with a flexible Tukey-lambda distribution for the innovations, we can extrapolate the high quantile levels by borrowing information from intermediate ones. Asymptotic properties for the proposed estimators are established. Simulation experiments are carried out to access the finite sample performance of the proposed methods, and an empirical example is presented to illustrate the usefulness of the new model.

#### Session 23CHI19: Recent Development in Extreme Value Theory

##### Estimation and Inference for Extreme Continuous Treatment Effects

Wei Huang<sup>1</sup>, ♦Shuo Li<sup>2</sup> and Liuhua Peng<sup>1</sup>

<sup>1</sup>The University of Melbourne

<sup>2</sup>Tianjin University of Finance and Economics

shuoli@tjufe.edu.cn

This paper studies estimation and inference for the treatment effect on deep tails of the potential outcome distributions corresponding to a continuously valued treatment, namely the extreme continuous treatment effect. We consider two measures for the tail characteristics: the quantile function and the tail mean function defined as the conditional mean beyond a quantile level. Then for a quantile level close to 1, we define the extreme quantile treatment effect (EQTE) and extreme average treatment effect (EATE), which are respectively the differences of the quantile and tail mean at different treatment statuses. We propose estimators for the EQTE and EATE based on tail approximations from the extreme value theory. Our limiting theory is for the EQTE and EATE processes indexed by a set of quantile levels and hence facilitates uniform inference for the EQTE and EATE over multiple tail levels. Simulations suggest that our method works well in finite samples and two empirical studies illustrate its practical merits.

## Abstracts

---

### Panel Quantile Regression for Extreme Risk

Yanxi Hou<sup>1</sup>, ♦Xuan Leng<sup>2</sup>, Liang Peng<sup>3</sup> and Yinggang Zhou<sup>2</sup>

<sup>1</sup>Fudan University

<sup>2</sup>Xiamen University

<sup>3</sup>Georgia State University  
xleng@xmu.edu.cn

Panel quantile regression models play an essential role in finance, insurance, and risk management applications. However, a direct application of panel regression for the extreme conditional quantiles may suffer from significant estimation errors due to data sparsity on the far tail. We introduce a two-stage method to predict extreme conditional quantiles over cross-sections. First, use panel quantile regression at a selected intermediate level, then extrapolate the intermediate level to an extreme level with extreme value theory. The combination of panel quantile regression at an intermediate level and extreme value theory relies on a set of second-order conditions for heteroscedastic extremes. We also propose a metric called *Average Absolute Relative Error* to evaluate the prediction performance of both intermediate and extreme conditional quantiles. Individual fixed effects in panel quantile regressions complicate the asymptotic analysis of the two-stage method and prediction metric. We demonstrate the finite sample performance of the extreme conditional quantile prediction compared to the direct use of panel quantile regression. Finally, we apply the two-stage method to the macroeconomic and housing price data and find strong evidence of housing bubbles and common economic factors.

### Simultaneous Confidence Bands for Conditional Value-at-Risk and Expected Shortfall

Shuo Li<sup>1</sup>, ♦Liuhsia Peng<sup>2</sup> and Xiaojun Song<sup>3</sup>

<sup>1</sup>Tianjin University of Finance and Economics

<sup>2</sup>The University of Melbourne

<sup>3</sup>Peking University  
liuhua.peng@unimelb.edu.au

Conditional value-at-risk (CVaR) and conditional expected shortfall (CES) are widely adopted risk measures which help monitor potential tail risk while adapting to evolving market information. In this paper, we propose an approach to construct simultaneous confidence bands (SCBs) for tail risk as measured by CVaR and CES, with the confidence bands uniformly valid for a set of tail levels. We consider one-sided tail risk (downside or upside tail risk) as well as relative tail risk (the ratio of upside to downside tail risk). A general class of location-scale models with heavy-tailed innovations is employed to filter out the return dynamics. Then, CVaR and CES are estimated with the aid of extreme value theory. In the asymptotic theory, we consider two scenarios: (i) the extreme scenario that allows for extrapolation beyond the range of the available data and (ii) the intermediate scenario that works exclusively in the case where the available data are adequate relative to the tail level. For finite-sample implementation, we propose a novel bootstrap procedure to circumvent the slow convergence rates of the SCBs as well as infeasibility of approximating the limiting distributions. A series of Monte Carlo simulations confirm that our approach works well in finite samples.

### Session 23CHI20: High Dimensional Modeling and Inference

#### Structure Learning of a Gaussian Amp Chain Graph via Acyclicity Constraints

♦Wei Zhou and Wei Zhong

Xiamen University  
zhouwei@stu.xmu.edu.cn

Chain graphs (CG) have been widely used to represent the undirected conditional dependence and directed causal relationships in one graphical model. It is of great challenges to learn the structure of a chain graph due to the combinational constraint of no semi-directed cycles and the existing methods are based on the independence tests, which lead to extensive computation cost. In this paper, we propose a novel constraint to characterize the acyclicity with the directed adjacency matrix and undirected precision matrix and cast the combinatorial optimization problem as a continuous one. The proposed method for the structural learning problem of CGs reduces the number of constraints from super-exponentially complexity to polynomial time. Particularly, we establish the theoretical guarantees for the consistency of the reconstructed CG when identifiable. Numerical studies demonstrate the efficiency of the proposed method. The transport network data for China's ports is analyzed to illustrate the relationships among the economic giants cities and clusters.

#### New Tests for High-Dimensional Two-Sample Mean Problems with Consideration of Correlation Structure

♦Songshan Yang<sup>1</sup>, Shurong Zheng<sup>2</sup> and Runze Li<sup>3</sup>

<sup>1</sup>Renmin University of China

<sup>2</sup>Northeast Normal University

<sup>3</sup>The Pennsylvania State University  
songshan.yang@gmail.com

This paper proposes a test statistic for two sample mean testing problems for high dimensional data by assuming the linear structure on high dimensional precision matrices. A new precision matrix estimation method considering its linear structure is first proposed, and the regularization method is implemented to select the true basis matrices that can further reduce the approximation error. Then the test statistic is constructed by imposing the estimation of the precision matrix. The proposed test is valid for both the low dimensional setting and high dimensional setting even if the dimension of the data is greater than the sample size. The limiting null distributions of the proposed test statistic under both null distribution and alternative distribution are derived. Extensive simulations are conducted for estimating the precision matrix and testing difference of the high dimensional mean vector. Simulation results show that the proposed estimation method enjoy low estimation error for the precision matrix and the regularization method is able to efficiently select the important basis matrix. The testing method performs well compared with the existing methods especially when the elements of the vector have unequal variances. A real data example is then provided to demonstrate the potential of the proposed method

#### Projective Independence Tests in High Dimensions: The Curses and the Cures

♦Yaowu Zhang<sup>1</sup> and Liping Zhu<sup>2</sup>

<sup>1</sup>Shanghai University of Finance and Economics

<sup>2</sup>Renmin University of China  
zhang.yaowu@mail.shufe.edu.cn

Testing independence between high dimensional random vectors is fundamentally different from testing independence between univariate random variables. Take the projection correlation as an example. It suffers from at least three issues. First, the complexity of estimating the projection correlation is of order  $O\{n^{3/2}(p+q)\}$ , where  $n$ ,  $p$  and  $q$  are the respective sample size and dimensions of the random vectors. This limits its usefulness substantially when  $n$  is large. Second, the asymptotic null distribution of the projection correlation test is rarely tractable. Therefore, random permutations are often

suggested to approximate the asymptotic null distribution. This further increases the complexity of implementing independence tests. Last, the power performance of the projection correlation test deteriorates in high dimensions. To address these issues, we improve the projection correlation through a modified weight function, which reduces the complexity of estimating the projection correlation to the order of  $O\{n^2(p+q)\}$ . We estimate the improved projection correlation with the U-statistic theory. More importantly, its asymptotic null distribution is standard normal, thanks to the high dimensions of random vectors. This expedites the implementation of independence tests substantially. To enhance power performance in high dimensions, we introduce a cross-validation procedure which incorporates feature screening with the projection correlation test.

#### **Large-Scale Two-Sample Comparison of Support Sets**

*Haoyu Geng<sup>1</sup>, Xiaolong Cui<sup>1</sup>, ♦Haojie Ren<sup>2</sup> and Changliang Zou<sup>1</sup>*

<sup>1</sup>Nankai University

<sup>2</sup>Shanghai Jiao Tong University

haojieren@sjtu.edu.cn

Two-sample multiple testing has a wide range of applications. Most of the literature considers simultaneous tests of equality of parameters. This talk takes a different perspective and investigates the null hypotheses that the two support sets are equal. This formulation of the testing problem is motivated by the fact that in many applications where the two parameter vectors being compared are both sparse, one might be more concerned about the detection of differential sparsity structures rather than the difference in parameter magnitudes. Focusing on this type of problems, we develop a general approach, which adapts the newly proposed symmetry data aggregation tool combined with a novel double thresholding (DT) filter. The DT filter first constructs a sequence of pairs of ranking statistics that fulfill global symmetry properties, and then chooses two data-driven thresholds along the ranking to simultaneously control the false discovery rate (FDR) and maximize the number of rejections. Several applications of the methodology are given including high-dimensional linear models and Gaussian graphical models. We show that the proposed method is able to asymptotically control the FDR under certain conditions. Numerical results confirm the effectiveness and robustness of DT in FDR control and detection ability in many settings.

#### **Session 23CHI21: Statistical Learning for Regression Analysis and Change Point Detection**

##### **Online Smooth Backfitting for Generalized Additive Models**

♦Ying Yang<sup>1</sup>, Fang Yao<sup>2</sup> and Peng Zhao<sup>3</sup>

<sup>1</sup>Academy of Mathematics and Systems Science, Chinese Academy of Sciences

<sup>2</sup>Peking University

<sup>3</sup>Jiangsu Normal University  
yangy2022@amss.ac.cn

We propose an online smoothing backfitting method for generalized additive models coupled the local linear estimation. The idea can be extended to the general nonlinear optimization problems. The strategy is to use an appropriate-order expansion to approximate the nonlinear equations and store the coefficients as sufficient statistics which can be updated in an online manner by the dynamic candidate bandwidth method. We investigate the statistical and algorithmic convergences of the proposed method. By defining the relative statistical efficiency and computational cost, we further establish a framework to characterize the trade-off between the estimation and

computation performances. Simulations and real data examples are provided to illustrate the proposed method and algorithm.

##### **Composite Smoothed Quantile Regression**

Xiaozhou Wang

East China Normal University

shirley\_wxz@126.com

Composite quantile regression (CQR) is an efficient method to estimate the parameters of the linear model with non-Gaussian random noise. The non-smoothness of CQR loss prevents many efficient methods from being used. In this paper, we propose the composite smoothed quantile regression model and investigate the inference problem for a large-scale dataset. The algorithm and theoretical properties are given. Extensive numerical experiments on both simulated and real data are conducted to demonstrate the good performance of the proposed estimator compared to some baselines.

##### **Detecting Multi-Threshold Effects Ofaccelerate Failure Time Model by Sample-Splitting Methods**

Chuang Wan

wanchuang@nankai.edu.cn

TBC

#### **Session 23CHI23: Statistics Process Control and Change-Point Analysis**

##### **Stochastic Integral Bootstrap for Statistics of Irregularly Spaced Spatial Data**

Shibin Zhang

Shanghai Normal University

zhang\_shibin@shnu.edu.cn

Resampling approaches to approximate the sampling distribution of the statistic constructed from irregularly spaced data are still far from well-developed. We propose a novel bootstrap method, the stochastic integral bootstrap (SIB), as a complement to the existing approaches. It uses a stochastic integral (SI) to approximate the sampling distribution of a statistic rather than generating a resample from the observation. Meanwhile, the variance of the SI is consistent to that of the statistic. The SI is constructed by integrating a block-constrained version of the statistic with respect to the standard Gaussian white noise, which provides a general class of resampling estimators based on irregularly spaced spatial data. The proposed SIB imitates the second-order dependence of multiple statistics very well. As a result, it can be used to approximate the sampling distribution of multivariate statistics achieving the joint asymptotic normality. In this talk, the SIB is applicable to various important statistics such as the sample mean, the sample variance, the auto-covariance estimator, the discrete Fourier transform and some test statistics for spatial white noise as well. In addition, simulation studies illustrate the finite sample performance of the SIB in comparison with some competitive counterparts for irregularly spaced spatial data.

##### **An Optimization Method for Monitoring Change-Point in Small Samples**

♦Dong Han<sup>1</sup>, Fugee Tsung<sup>2</sup> and Jinguo Xian<sup>1</sup>

<sup>1</sup>Shanghai Jiao Tong University

<sup>2</sup>Hong Kong University of Science and Technology  
donghan@sjtu.edu.cn

This talk will present a method to construct an optimal control chart to detect a change in a sequence of finite or even small samples with the unknown change-point and the unknown post-change probability distribution. We prove that the proposed optimal control chart

## Abstracts

---

has the smallest average value of broadly defined detection delay with a false alarm rate not less than a given value. The power of method is illustrated by numerical simulations of three optimized control charts, Shewhart, EWMA and CUSUM charts, and a real data example.

### Surface Temperature Monitoring in Liver Procurement via Functional Variance Change-Point Analysis

♦Zhenguo Gao<sup>1</sup>, Pang Du<sup>2</sup>, Ran Jin<sup>2</sup> and John Robberson<sup>2</sup>

<sup>1</sup>Shanghai Jiao Tong University

<sup>2</sup>Virginia Tech

gaozhenguo3@126.com

Liver procurement experiments with surface-temperature monitoring motivated Gao et al. (J. Amer. Statist. Assoc. 114 (2019) 773–781) to develop a variance change-point detection method under a smoothly-changing mean trend. However, the spotwise change points yielded from their method do not offer immediate information to surgeons since an organ is often transplanted as a whole or in part. We develop a new practical method that can analyze a defined portion of the organ surface at a time. It also provides a novel addition to the developing field of functional data monitoring. Furthermore, numerical challenge emerges for simultaneously modeling the variance functions of 2D locations and the mean function of location and time. The respective sample sizes in the scales of 10,000 and 1,000,000 for modeling these functions make standard spline estimation too costly to be useful. We introduce a multistage subsampling strategy with steps educated by quickly-computable preliminary statistical measures. Extensive simulations show that the new method can efficiently reduce the computational cost and provide reasonable parameter estimates. Application of the new method to our liver surface temperature monitoring data shows its effectiveness in providing accurate status change information for a selected portion of the organ in the experiment.

### Robust Online Detection in Serially Correlated Directed Network

♦Miaomiao Yu<sup>1</sup>, Yuhao Zhou<sup>2</sup> and Fugee Tsung<sup>3</sup>

<sup>1</sup>East China Normal University

<sup>2</sup>University of North Carolina at Chapel Hill

<sup>3</sup>Hong Kong University of Science and Technology  
mmyu@ecnu.edu.cn

As the complexity of production processes increases, the diversity of data types drives the development of network monitoring technology. This paper mainly focuses on an online algorithm to detect serially correlated directed networks robustly and sensitively. First, we consider a transition probability matrix to resolve the double correlation of primary data. Further, since the sum of each row of the transition probability matrix is one, it standardizes the data, facilitating subsequent modeling. Then we extend the spring length based method to the multivariate case and propose an adaptive cumulative sum (CUSUM) control chart on the strength of a weighted statistic to monitor directed networks. This novel approach assumes only that the process observation is associated with nearby points without any parametric time series model, which is in line with reality. Simulation results and a real example from metro transportation demonstrate the superiority of our design.

### Session 23CHI24: Recent Developments in the Analysis of Complex Lifetime Data

#### Differential Equation-Assisted Local Polynomial Regression

W. John Braun

UBC

john.braun@ubc.ca

Local polynomial regression function estimation can often have better bias properties than local constant estimation, but the increase in degree can lead to instability. In this talk, we explore the situation where information about the first and/or second derivative of the regression can be used to improve accuracy while retaining stability. By exploiting the differential equation appropriately, a procedure more in line with local constant regression can be employed, partially circumventing missing data problems. This technique is illustrated with a problem involving growth curves.

#### Covariate Balancing with Measurement Error

Xialing Wen and ♦Ying Yan

Sun Yat-sen University

yanying7@mail.sysu.edu.cn

In the past decade, there is an emerging literature on developing covariate balancing methods among statisticians and applied researchers, where covariate balance is directly incorporated in the estimation procedure. It has been well documented that covariate balancing is superior to propensity score weighting in many circumstances. The validity of covariate balancing requires implicitly that all covariates are accurately measured. Measurement error is ubiquitous in real studies, but there is a lack of understanding of the role of measurement error in covariate balancing. In this talk, we systematically study the impact of measurement error on covariate balancing methods. In theory, we show that naive covariate balancing that ignores measurement error results in biased causal effect estimation and poor balancing performance. We then propose a class of measurement error correction strategies that successfully remove measurement error bias and achieve valid causal conclusions. Finally, we apply the proposed methods in simulation studies and the analysis of a lifetime data set.

#### Variable Selection for Recurrent Event Model with Covariates Subject to Measurement Error

♦Kaida Cai<sup>1</sup>, Hua Shen<sup>2</sup> and Xuewen Lu<sup>2</sup>

<sup>1</sup>Southeast University

<sup>2</sup>University of Calgary  
caikaida@seu.edu.cn

This article focuses on variable selection for recurrent event model when covariates are measured with errors. We propose a variable selection and estimation method to select important covariates, estimate the corresponding parameters and adjust for the effect of measurement error simultaneously. In the simulation study, we compare the proposed procedures with the methods for analyzing recurrent event data without considering variable selection or addressing the measurement error. The results of the simulation study show that the proposed procedures outperform the methods ignoring the measurement error issues or variable selection needs, especially in removing the unimportant error-prone covariates and estimating the corresponding parameters of important covariates. The numerical results also show that the bigger measurement error has greater negative impact on the variable selection results. We further illustrate the proposed procedures by an application.

#### Chair and Discussant

Hua Shen

University of Calgary  
hua.shen@ucalgary.ca

Not applicable

## Session 23CHI27: Stochastic Modeling and Inference of Epidemiological and Industrial Data with Complex Structures

### Statistical Inference of Multi-State Transition Model for Longitudinal Data with Measurement Error and Heterogeneity

♦*Jing Guan and Jiajie Qin*

Tianjin University  
guanjing@tju.edu.cn

Multi-state transition model is typically used to analyze longitudinal data in medicine and sociology. Moreover, variables in longitudinal studies usually are errorprone, and random effects are heterogeneous, which will result in biased estimates of the interest parameters. This paper is intended to estimate the parameters of multi-state transition model for longitudinal data with measurement error and heterogeneous random effects, and futher consider the covariate related to covariance matrix of random effects is also error-prone while the covariate in transition model is error-prone. We propose a pseudo-likelihood method based on Monte Carlo expectation maximization algorithm (MCEM) and bayesian method based on Markov Chain Monte Carlo to infer and calculate estimates and model the covariance matrix of random effects through the modified Cholesky ecomposition. Meanwhile, asymptotic property are obtained and the finite sample performance of the proposed method is verified by simulation, which is well in terms of Bias, RMSE and coverage rate of confidence intervals. In addition, the proposed method is applied to the MFUS data set.

### Bi-Level Selection with the Sparse Group Bar for Multivariate Interval- Censored Data

♦*Xuewen Lu and Fatemeh Mahmoudi*

University of Calgary  
xlu@ucalgary.ca

Multivariate interval-censored data arise when individuals in the study are at the risk of experiencing multiple events of interest and the onset time of each event is not observed precisely, but it is known to lie in a time window between two examinations. Regularized regression is a popular approach in the framework of variable selection problems. In this regard, we propose a new composite penalty to incorporate the oracle property and grouping effect property of the broken adaptive ridge (BAR) regression and a group bridge (GB) penalty, in contrast to the existing sparse group lasso, we term it as the sparse group BAR. This leads to a new bi-level variable selection method that selects variables at both individual and group levels. We propose a novel EM based algorithm that encompasses data augmentation and iterative least square approximation to solve the optimization problem. Simulation studies and data analysis are conducted to demonstrate the performance of the proposed method.

### Euler's Number, Euler Numbers, and Eulerian Numbers

*James C. Fu<sup>1</sup>, Wan-Chen Lee and ♦Hsing-Ming Chang<sup>2</sup>*

<sup>1</sup>University of Manitoba

<sup>2</sup>National Cheng Kung University  
nckuhmchang@mail.ncku.edu.tw

Everyone knows Euler's number ( $e = 2.718\dots$ ). We present in this talk a method, using non-homogeneous finite Markov chains, to obtain the exact distribution of the number of peaks associated with a random permutation of  $1, 2, \dots, n$ . Not only the number of alternating permutations in the group of all permutations generated by  $1, 2, \dots, n$  for given  $n$  can be obtained, along the way, its connection to the number of descents (falls) is established through the distribution of peaks.

## A Censored Quantile Transformation Model for Alzheimer's Disease Data with Multiple Functional Covariates

*Maozai Tian*

Renmin University of China  
mztian@ruc.edu.cn

Alzheimer's disease (AD) is a progressive disease that starts from mild cognitive impairment and may finally lead to memory loss. Therefore, it is critical to explore the risk factors for the conversion time to AD. However, previous studies in this context mainly focus on modeling single functional covariate, leaving out many important information. In our motivating AD data, both the left and right hippocampal radial distance curves are potentially significant, but they are correlated with each other, which may cause serious multicollinearity in the regression. Thus, we propose a multivariate functional censored quantile regression model based on the Box-Cox transformation to analyze the right-censored data. The multiple functional predictors are jointly handled by the multivariate functional principal component analysis. By introducing dynamic power transformations, the proposed method relaxes the global linear assumption imposed by many existing studies and enjoys more flexibility in the framework of censored quantile regression. Based on the martingale theory, uniform consistency and weak convergence are established as a process of quantile levels. Simulation studies demonstrate the outperformance of the proposed method. Real data analysis shows the importance of both left and right hippocampal radial distance curves for predicting the conversion time to AD in different quantile

## Session 23CHI79: Novel Bayesian Methods for Complex Dataand their Applications

### A Bayesian Joint Model of Longitudinal and Time-to-Event Data for the Aids Treatment Effectiveness Evaluation

♦*Tao Wang and Yinxiang Zhang*  
wtaokm@263.net

In the traditional joint models of a longitudinal and time-to-event data, a linear mixed effects model assuming normal random errors is used to model the longitudinal process. However, in many HIV clinical data especially in Chinese HIV clinical data, the normality assumption is violated and the linear mixed model is not an appropriate sub-model in the joint models. In addition the quantiles of the longitudinal process is the main concern of the clinician. So we adopt a fully Bayesian version that the linear quantile mixed model for the longitudinal process in the joint model, implemented via Bayesian Method and Markov chain Monte Carlo (MCMC) methods to estimate the parameters in our joint model . We use the approach to jointly model the longitudinal and survival data from an AIDS clinical trial comparing two treatments to illustrate the good performance of our method.

### Statistical Inference for Copula-Based Dependent Competing Risks Model with Step-Stress Accelerated Life Test

*Wenhao Gui*

Beijing Jiaotong University  
whgui@bjtu.edu.cn

In this talk, a dependent competing risks model is considered and investigated by utilizing the copula approach which flexibly constructs dependent relationships between marginal distributions of several competing risk factors, and the closeness of the dependency can be determined by the copula parameters. Samples are collected from the step-stress accelerated life test combined with generalized

## Abstracts

progressive hybrid censoring, where the test duration is controlled within acceptable limits and a sufficient number of samples are guaranteed. From a classical frequency perspective, the maximum likelihood estimates are derived, then the relevant confidence intervals are provided based on the Fisher information matrix. Furthermore, the bias-corrected accelerated bootstrap method is employed to obtain the interval estimation of parameters. Under the Bayesian framework, Bayesian point estimates and the corresponding credible intervals are also explored, and their results are achieved by the Markov Chain Monte Carlo technique. Finally, numerical simulation experiments and real data analysis are developed to present the constructed model and the performance of the above-mentioned methods.

### Bayesian Jackknife Empirical Likelihood for the Error Variance in Linear Regression Models

♦Hongyan Jiang<sup>1</sup> and Yichuan Zhao<sup>2</sup>

<sup>1</sup>Department of Mathematics and Physics, Huaiyin Institute of Technology, Huaian, People's Republic of China

<sup>2</sup>Department of Mathematics and Statistics, Georgia State University, Atlanta, GA, United States  
hyitjhy@hotmail.com

Variance estimation is fundamental in the statistical inference. Due to the nonlinearity of the variance estimator, Lin et al. proposed the jackknife empirical likelihood method for the error variance in a linear regression model. However, people may have some prior information about the error variance. In this article, we propose the Bayesian jackknife empirical likelihood (BJEL) for the error variance in a linear regression model. The validity of the proposed method is verified, and the asymptotic normal properties for the BJEL are also established. A simulation study shows that the new approach for the small sample performs better than its frequentist counterpart. Two real data sets are also used to illustrate the proposed methods.

### Changepoint Joint Modeling of High-Dimensional Longitudinal and Survival Data: a Bayesian Approach

Yangxin Huang

University of South Florida  
yhuang@usf.edu

Joint modeling of longitudinal-survival data is an active area of statistical research. Although it is a common practice to analyze complex longitudinal data using NLME models, the following issues may standout: (i) In practice, the profile of subject's longitudinal responses may follow a "broken-sticklike trajectory, indicating multiple phases of increase and/or decline in responses. To estimate random change-points, NLME models become a challenge due to complicated model structure. (ii) Many studies are often to collect multivariate longitudinal responses which may be significantly correlated, ignoring their correlation may lead to bias and reduce efficiency in estimation. Moreover, the time-to-event may be dependent on the multivariate longitudinal measures. (iii) Missing observations in the longitudinal responses are often encountered; the missing data are likely to be informative. Under Bayesian framework we consider a piecewise joint model for multivariate longitudinal and time-to-event data to estimate change rates of longitudinal trajectory patterns and timing of change-point which may be critical indicators to quantify the effect of longitudinal profile on the risk of an event. The proposed models and method are applied to analyze a longitudinal data set arising from a diabetes study. Simulation studies are conducted to assess the performance of the proposed method.

## Session 23CHI1: Sufficient Dimension Reduction and Beyond

### Deep Nonlinear Sufficient Dimension Reduction

Fengyin Chen<sup>1</sup>, Yuling Jiao<sup>2</sup>, Rui Qiu<sup>1</sup> and ♦Zhou Yu<sup>1</sup>

<sup>1</sup>East China Normal University

<sup>2</sup>Wuhan University  
zyu@stat.ecnu.edu.cn

Linear sufficient dimension reduction, as exemplified by sliced inverse regression, has seen substantial development in the past thirty years. However, with the advent of more complex scenarios, nonlinear dimension reduction has become a more general topic that gains considerable interests recently. This article introduces a novel method for nonlinear sufficient dimension reduction, utilizing the generalized martingale difference divergence measure in conjunction with deep neural networks. The optimal value of the objective function is shown to be unbiased at the general level of -fields. And the optimization scheme based on the fascinating deep neural networks is more efficient and flexible than the classical eigen-decomposition of linear operators. Moreover, we systematically investigate the slow rate and fast rate for the estimation error based on advanced U-process theory. Remarkably, the fast rate is nearly minimax optimal. The effectiveness of the deep nonlinear sufficient dimension reduction method is demonstrated through simulations and real data analysis.

### Dimension Reduction for Frechet Regression

Qi Zhang, Lingzhou Xue and ♦Bing Li

Pennsylvania State University  
bxl9@psu.edu

With the rapid development of data collection techniques, complex data objects that are not in the Euclidean space are frequently encountered in new statistical applications. Frechet regression model (Peterson & Muller 2019) provides a promising framework for regression analysis with metric space-valued responses. In this paper, we introduce a flexible sufficient dimension reduction (SDR) method for Frechet regression to achieve two purposes: to mitigate the curse of dimensionality caused by high-dimensional predictors, and to provide a visual inspection tool for Frechet regression. Our approach is flexible enough to turn any existing SDR method for Euclidean  $(X, Y)$  into one for Euclidean  $X$  and metric space-valued  $Y$ . We established the consistency and asymptotic convergence rate of the proposed methods. The finite-sample performance of the proposed methods is illustrated through simulation studies for several commonly encountered metric spaces that include Wasserstein space, the space of symmetric positive definite matrices, and the sphere. We illustrated the data visualization aspect of our method by the human mortality distribution data from the United Nations Databases.

### Sliced Inverse Regression with Large Structural Dimension

Dongming Huang<sup>1</sup>, Songtao Tian<sup>2</sup> and ♦Qian Lin<sup>2</sup>

<sup>1</sup>NUS

<sup>2</sup>Tsinghua University  
qianlin@tsinghua.edu.cn

The central space of a joint distribution  $(X, Y)$  is the minimal subspace  $\mathcal{S}$  such that  $Y \perp\!\!\!\perp X | P_{\mathcal{S}}X$  where  $P_{\mathcal{S}}$  is the projection onto  $\mathcal{S}$ . Sliced inverse regression (SIR), one of the most popular methods for estimating the central space, often performs poorly when the structural dimension  $d = \dim(\mathcal{S})$  is large (e.g.,  $\geq 5$ ). In this paper, we demonstrate that the generalized signal-noise-ratio (gSNR) tends to be extremely small for a general multiple-index model when  $d$  is large. Then we determine the minimax rate for

estimating the central space over a large class of high dimensional distributions with large structural dimension  $d$  (i.e., there is no constant upper bound on  $d$ ) in the low gSNR regime. This result not only extends the existing minimax rate results for estimating the central space of distributions with fixed  $d$  to that with large  $d$ , but also clarifies that the degradation in SIR performance is caused by the decay of signal strength. The technical tools developed here might be of independent interest for studying other central space estimation methods.

#### A Unified Generalization of Inverse Regression via Adaptive Column Selection

*Yin Jin and ♦Wei Luo*

Zhejiang University  
wluostat@outlook.com

Higher-order inverse regression methods are commonly known as more powerful sufficient dimension reduction (SDR) methods than the popularly used sliced inverse regression (SIR) in the population level. However, due to the convention of essentially conducting singular value decomposition on the ambient candidate matrices, these methods suffer from the excessive number of parameters in the sample level and have not been systematically generalized under the high-dimensional settings like SIR. In this paper, we break the convention of using the ambient candidate matrices in these methods, and instead apply a novel column-selection strategy on their candidate matrices that substantially lowers down the working number of parameters to being comparable with SIR. Then, for the first time of the literature, we generalize the higher-order inverse regression methods, as well as their ensembles, towards sparsity under the high-dimensional settings in a uniform manner. The dimension of the predictor is allowed to diverge with the sample size in nearly an exponential order, and no additional restrictions are imposed on the data other than those commonly seen in the high-dimensional literature. For completeness of theory, we also study the column-selection strategy towards the estimation efficiency under the conventional low-dimensional settings. These results are illustrated by simulations

#### Session 23CHI10: Novel Bayesian Adaptive Clinical Trial Designs and Methods for Precision Medicine

##### Sam: Self-Adapting Mixture Prior to Dynamically Borrow Information from Historical Data in Clinical Trials

*Peng Yang<sup>1</sup>, Yuansong Zhao<sup>2</sup>, Lei Nie<sup>3</sup>, Jonathon Vallejo<sup>3</sup> and ♦Ying Yuan<sup>4</sup>*

<sup>1</sup>Rice University

<sup>2</sup>The University of Texas Health Science Center

<sup>3</sup>FDA

<sup>4</sup>The University of Texas MD Anderson Cancer Center  
yyuan@mdanderson.org

Mixture priors provide an intuitive way to incorporate historical data while accounting for potential prior-data conflict by combining an informative prior with a non-informative prior. However, specifying the mixing weight for each component remains a crucial challenge. Ideally, the mixing weight should reflect the degree of prior-data conflict, which is often unknown beforehand, posing a significant obstacle to the application and acceptance of mixture priors. To address this challenge, we introduce self-adapting mixture (SAM) priors that determine the mixing weight using likelihood ratio test statistics. SAM priors are data-driven and self-adapting, favoring the informative (non-informative) prior component when

there is little (substantial) evidence of prior-data conflict. Consequently, SAM priors achieve dynamic information borrowing. We demonstrate that SAM priors exhibit desirable properties in both finite and large samples and achieve information-borrowing consistency. Moreover, SAM priors are easy to compute, data-driven, and calibration-free, mitigating the risk of data dredging. Numerical studies show that SAM priors outperform existing methods in adopting prior-data conflicts effectively. We developed an R package and web application that are freely available to facilitate the use of SAM priors.

##### Adaptive Promising Zone Design for Cancer Immunotherapy with Heterogenous Delayed Treatment Effects

*Bosheng Li<sup>1</sup>, Fangrong Yan<sup>1</sup> and ♦Depeng Jiang<sup>2</sup>*

<sup>1</sup>China Pharmaceutical University

<sup>2</sup>University of Manitoba

Depeng.Jiang@umanitoba.ca

The indirect mechanism of immunotherapy for cancer might lead to the delayed treatment effect and the delay times are often heterogeneous among patients. Using the conventional log-rank test and Cox regression for the immunotherapy trial design and analysis could result in a sever loss of power and biased estimation of treatment effect. In this paper, we proposed a promising zone design with an interim analysis. We first used the interim data to re-estimate the parameters in the survival function and the distribution of delayed time. Then we calculated the conditional power and defined the promising zones. We conducted simulation studies to illustrate the merits of our proposed promising zone design over the fixed sample design and the group sequential design. The results indicate that our proposed promising zone design improved the conditional power remarkably over the fixed sample design. Our proposed method can control Type I errors very well without sacrificing conditional power compared with the group sequential design. Our proposed design method may be attractive to some small biotech companies because it allows a small initial sample size and then might attract venture capital investments to increase the sample size based on the promising interim result. We will also

##### A Generalized Phase 1-2-3 Design Integrating Dose Optimization with Confirmatory Treatment Comparison

*Yong Zang*

Indiana University

zangy@iu.edu

A generalized phase 1-2-3 design, Gen 1-2-3, that includes all phases of clinical treatment evaluation is proposed. It uses phase 1-2 criteria to identify a set of candidate doses rather than one dose. In an intermediate stage between phase 1-2 and phase 3, it randomizes additional patients fairly among the candidate doses and an active control treatment arm and uses survival time data to select an optimal dose. It makes a Go / No Go decision of whether or not to conduct phase 3 based on a predictive probability that the optimal dose will provide a substantive improvement over the control in survival time. A simulation study shows that the Gen 1-2-3 design has desirable operating characteristics compared to the CT design and two conventional designs.

#### Session 23CHI11: Advanced Biostatistics and Bioinformatics Approaches for Medical Research

##### The Identifiability of Copula Models for Dependent Competing Risks Data with Exponentially Distributed Margins

*Antai Wang*

## Abstracts

---

New Jersey Institute of Technology  
aw224@njit.edu

We prove the identifiability property of Archimedean copula models for dependent competing risks data when at least one of the failure times is exponentially distributed. With this property, it becomes possible to quantify the dependence between competing events based on exponentially distributed dependent censored data. We demonstrate our estimation procedure using simulation studies and in an application to survival data.

### **Adaptive Bayesian Phase i Clinical Trial Designs for Estimating the Maximum Tolerated Doses for Two Drugs While Fully Utilizing all Toxicity Information**

Zhengjia Chen

University of Illinois at Chicago  
znchen@uic.edu

The combined treatments with multiple drugs are very common in the contemporary medicine, especially for medical oncology. Therefore, in this talk, I will introduce our newly developed Bayesian adaptive Phase I clinical trial design entitled Escalation with Overdoing Control using Normalized Equivalent Toxicity Score for estimating maximum tolerated dose (MTD) contour of two drug COMbination (EWOC-NETS-COM) used for oncology trials. The normalized equivalent toxicity score (NETS) as the primary endpoint of clinical trial is assumed to follow quasi-Bernoulli distribution and treated as quasi-continuous random variable in the logistic linear regression model which is used to describe the relationship between the doses of the two agents and the toxicity response. Four parameters in the dose-toxicity model were re-parameterized to parameters with explicit clinical meanings to describe the association between NETS and doses of two agents. Non-informative priors were used and Markov Chain Monte Carlo was employed to update the posteriors of the 4 parameters in dose-toxicity model. Extensive simulations were conducted to evaluate the safety, trial efficiency, and MTD estimation accuracy of EWOC-NETS-COM under different scenarios, using the EWOC as reference. The results demonstrated that EWOC-NETS-COM not only efficiently estimates MTD contour of multiple drugs, but also provides better trial efficiency

### **Developing Machine Learning Models to Improve Fracture Prediction using Genomic and Phenotypic Data in Large-Scale Observational Studies**

Qing Wu

The Ohio State University  
qing.wu@osumc.edu

This study developed fracture prediction models using machine learning and genomic data from large-scale observational studies. Many machine learning models, including random forest, gradient boosting, neural network, and logistic regression, were used to predict major osteoporotic fractures using GRS and other risk factors as predictors. In the Osteoporotic Fractures in Men cohort Study ( $n = 5130$ ), the best model was gradient boosting with an AUC of 0.71 and an accuracy of 0.88, and the GRS was ranked as the 7th most important variable. In the Women's Health Initiative study ( $n=25,772$ ), the XGBoost emerged as the top-performing model, achieving an accuracy of 91.3% and a weighted F1 score of 0.904; GRS was ranked as the 5th most important predictor. These studies suggest that incorporating genetic profiling and the gradient-boosting approach can improve fracture prediction.

### **using Auxiliary Information in Probability Survey Data to Improve Pseudo-Weighting in Non-Probability Samples: a Copula**

## **Model Approach**

Tingyu Zhu, ♦Lan Xue and Ginny Lesser

Oregon State University  
xuel@stat.oregonstate.edu

While probability sampling has been considered the gold standard of survey methods, nonprobability sampling such as online opt-in surveys is increasingly popular due to its convenience and low cost. However, nonprobability samples can lead to biased estimates due to the unknown nature of the underlying selection mechanism. In this paper, we propose both parametric and semi-parametric approaches to integrate probability and nonprobability samples based on common ancillary variables observed in both samples. In the parametric approach, the joint distribution of ancillary variables is assumed to follow the latent Gaussian copula model, which is flexible to accommodate both discrete and continuous variables. In contrast, the semi-parametric approach requires no assumptions about the distribution of ancillary variables. The proposed method is evaluated in the context of estimating the population mean where the estimators are constructed based on the probability sample, the non-probability sample with estimated inclusion probabilities, and the combined samples. Our simulation results show that the proposed method is able to correct the selection bias in the nonprobability sample. A real data application is provided to illustrate the practical use of the proposed method.

## **Session 23CHI12: Topics on Latent Variable Models and Minorisation-Maximisation Algorithm**

### **Sufficient Dimension Reductions under the Mixtures of Multivariate Elliptical Distributions**

Wenjuan Li, Hongming Pei, Ali Jiang and ♦Fei Chen

Yunnan University of Finance and Economics  
chen\_fei\_77@hotmail.com

In the sufficient dimension reduction (SDR), many methods depend on some assumptions on the distribution of predictor vector, such as the linear design condition (L.D.C.), the assumption of constant conditional variance, and so on. The mixture distributions emerge frequently in practice, but they may not satisfy the above assumptions. In this article, a general framework is proposed to extend various SDR methods to the cases where the predictor vector follows the mixture elliptical distributions, together with the asymptotic property for the consistency of the kernel matrix estimators. For illustration, the extensions of several classical SDR approaches under the proposed framework are detailed. A resampling algorithm is also introduced into the proposed framework to provide another option of the kernel matrix estimate. Moreover, a method to estimate the structural dimension is given, together with a procedure to check an assumption called homogeneity. The proposed methodology is illustrated by simulated and real examples.

### **Earthquake Parametric Insurance with Bayesian Spatial Quantile Regression**

Jeffrey Pai<sup>1</sup>, ♦Yunxian Li<sup>2</sup>, Aijun Yang<sup>3</sup> and Chenxu Li<sup>2</sup>

<sup>1</sup>University of Manitoba

<sup>2</sup>Yunnan University of Finance and Economics

<sup>3</sup>Nanjing Forest University  
liyunxian71@hotmail.com

With its transparent and fast claims payment, parametric insurance has been widely used to insure nature-related risks such as earthquakes, floods, and hurricanes. In 2014, earthquake parametric insurance was introduced to provide coverage for earthquake losses

occurred in Yunnan Province of China. However, as a main limitation of parametric insurance, basis risk is inevitable. In this paper, a Bayesian spatial quantile regression model is proposed to reduce the basis risk of earthquake parametric insurance. The effect of earthquake hazard, risk exposure, and vulnerability on economic loss are analyzed and considered in the quantile regression model. Since risk exposure and vulnerability at the epicenter cannot be observed, they will be treated as latent variables in the quantile regression model. Bayesian approaches are applied, and spatial correlation is considered to construct the prior distributions for the latent variables. Earthquake losses in Yunnan Province from 1992 to 2019 are collected and analyzed by the proposed model and methods. The payment mechanism and the corresponding premiums of 16 regions in Yunnan Province are then calculated. The results show that the loss ratio is more reasonable than the current earthquake insurance, and the basis risk is then reduced.

#### **Proportional Inverse Gaussian Distribution: a New Tool for Analysing Continuous Proportional Data**

♦*Pengyi Liu<sup>1</sup>, Guo-Liang Tian<sup>2</sup>, Kam Chuen Yuen<sup>3</sup>, Chi Zhang<sup>4</sup> and Man-Lai Tang<sup>5</sup>*

<sup>1</sup>Yunnan University of Finance and Economics

<sup>2</sup>Southern University of Science and Technology

<sup>3</sup>The University of Hong Kong

<sup>4</sup>Shenzhen University

<sup>5</sup>Brunel University London  
liupy@ynufe.edu.cn

Outcomes in the form of rates, fractions, proportions and percentages often appear in various fields. Existing beta and simplex distributions are frequently unable to exhibit satisfactory performances in fitting such continuous data. This paper aims to develop the normalized inverse Gaussian (N-IG) distribution proposed by Lijoi, Mena & Prünster (2005) as a new tool for analysing continuous proportional data in (0; 1) and renames the NIG as proportional inverse Gaussian (PIG) distribution. Our main contributions include: (1) To overcome the difficulty of an integral in the PIG density function, we propose a novel minorisation – maximisation (MM) algorithm via the continuous version of Jensen's inequality to calculate the maximum likelihood estimates of the parameters in the PIG distribution; (2) We also develop an MM algorithm aided by the gradient descent algorithm for the PIG regression model, which allows us to explore the relationship between a set of covariates with the mean parameter; (3) Both the comparative studies and the real data analyses show that the PIG distribution is better when comparing with the beta and simplex distributions in terms of the AIC, the Cramér–von Mises and the Kolmogorov–Smirnov test. In addition, bootstrap confidence intervals and testing hypothesis on the symmetry of the

#### **Session 23CHI13: New Developments in Statistical Detection**

##### **Modeling Autoregressive Conditional Regional Extremes with Application to Solar Flare Detection**

*Jili Wang<sup>1</sup> and ♦Zhengjun Zhang<sup>2</sup>*

<sup>1</sup>University of Wisconsin

<sup>2</sup>University of Chinese Academy of Sciences  
zhangzhengjun@ucas.ac.cn

This paper studies big data streams with regional-temporal extreme event (REE) structures and solar flare detection. An autoregressive conditional Fréchet model with time-varying parameters for

regional and its adjacent regional extremes (ACRAE) is proposed. The ACRAE model can quickly detect rare REEs (i.e., solar flares) in big data streams and predict solar activity. The ACRAE model, with some mild regularity conditions, is proved to be stationary and ergodic. The parameter estimators are derived through the conditional maximum likelihood method. The consistency and asymptotic normality of the estimators are established. Simulations are used to demonstrate the efficiency of the proposed parameter estimators. In real solar flare detection, with the new dynamic extreme value modeling, the occurrence and climax of solar activity can be detected earlier than existing algorithms. The empirical study shows that the ACRAE model outperforms the existing detection algorithms with sampling strategies.

##### **A Fdr –based Method for Monitoring High Dimensional Data Streams**

♦*Dequan Qi and Nan Liang*

School of Mathematics and Statistics, Changchun University of Science and Technology, Changchun Jilin  
qidequan@163.com

Based on a novel one-sided multivariate exponentially weighted moving average (MEWMA) chart, we propose a false discovery rate-adjusted scheme to on-line monitor high-dimensional data streams.

##### **Two Robust Multivariate Exponentially Weighted Moving Average Charts to Facilitate Distinctive Product Quality Features Assessment**

♦*Zhi Song<sup>1</sup>, Amitava Mukherjee<sup>2</sup>, Peihua Qiu<sup>3</sup> and Maoyuan Zhou<sup>4</sup>*

<sup>1</sup>Shenyang Agricultural University

<sup>2</sup>XLRI-Xavier School of Management

<sup>3</sup>University of Florida

<sup>4</sup>Civil Aviation University of China  
zhisong@syau.edu.cn

This paper offers new multivariate statistical process monitoring schemes to study the process shift supported by a distinctive product quality features assessment. The proposed procedures help determine whether one or more features have undergone a shift or whether the shift is in the dependence structure but not in the study variables. We use the marginal distributions and pseudo copula observations to effectively apply feature-wise rank-based Lepage and Cucconi tests. The rank-based statistics induce nonparametric properties to the proposed chart. Therefore, its in-control performance is highly robust and nearly distribution-free. It is shown that the new chart gives better results for detecting scale shifts in one or more quality variables than some representative existing nonparametric charts. Some Monte-Carlo simulation studies have been performed to establish the effectiveness of the new charting scheme. Two real applications are considered to illustrate the use of the proposed schemes in manufacturing and automobile quality online complaints. Some encouraging product feature assessment properties are observed.

##### **Enhancing Modeling and Monitoring Strategy for Directed Count-Weighted Networks**

♦*Chunjie Wu, Jinhua Qin and Wendong Li*

Shanghai University of Finance and Economics  
wumaths@mail.shufe.edu.cn

Network science has emerged as a prominent field of study, with applications in diverse domains such as transportation, social networks, and biological systems. In many applications, it is of great interest to determine whether and when the structure of a network

## Abstracts

---

changes over time. Network monitoring is a field of study dedicated to identifying such changes. The issue of statistical process monitoring for directed and weighted networks has received little attention in the literature. With this in mind, we adopt the gravity model to capture the directed and weighted network formation mechanism and proposed a new network monitoring methodology. With the node intensity parameter incorporated into the probabilistic model, the weight and direction of the network edges was characterized in a more intuitive manner. An efficient and effective estimation procedure was suggested as well. In order to utilize the samples in a more rational and effective way, we proposed to adopt the weighted likelihood ratio estimation to get an online parameter estimation, based on which a WLR-EWMA control chart was put forward. Finally, the proposed monitoring methodologies are illustrated through both a simulation study and a realdata application, showing their inherent advantages and effectiveness.

### Session 23CHI141: Fator Modelling and Data Analysis in Large-Scale Datasets

#### Factor Modelling for High-Dimensional Functional Time Series

*Shaojun Guo*

Renmin University of China  
sjguo@ruc.edu.cn

Many economic and scientific problems involve the analysis of high-dimensional functional time series, where the number of functional variables ( $p$ ) diverges as the number of serially dependent observations ( $n$ ) increases. In this paper, we present a novel functional factor model for high-dimensional functional time series that maintains and makes use of the functional and dynamic structure to achieve great dimension reduction and find the latent factor structure. To estimate the number of functional factors and the factor loadings, we propose a fully functional estimation procedure based on an eigenanalysis for a nonnegative definite matrix. Our proposal involves a weight matrix to improve the estimation efficiency and tackle the issue of heterogeneity, the rationality of which is illustrated by formulating the estimation from a novel regression perspective. Asymptotic properties of the proposed method are studied when  $p$  diverges at some polynomial rate as  $n$  increases. To provide a parsimonious model and enhance interpretability for near-zero factor loadings, we impose sparsity assumptions on the factor loading space and then develop a regularized estimation procedure with theoretical guarantees when  $p$  grows exponentially fast relative to  $n$ . Finally, we demonstrate that our proposed estimators significantly outperform the competing methods through both simulations and applications.

#### Network Gradient Descent Algorithm for Decentralized Federated Learning

♦*Shuyuan Wu<sup>1</sup>, Danyang Huang<sup>2</sup> and Hansheng Wang<sup>1</sup>*

<sup>1</sup>Peking University

<sup>2</sup>Renmin University of China  
shuyuan.w@pku.edu.cn

We study a fully decentralized federated learning algorithm, which is a novel gradient descent algorithm executed on a communication-based network. For convenience, we refer to it as a network gradient descent (NGD) method. In the NGD method, only statistics (e.g., parameter estimates) need to be communicated, minimizing the risk of privacy. Meanwhile, different clients communicate with each other directly according to a carefully designed network structure without a central master. This greatly enhances the reliability

of the entire algorithm. Those nice properties inspire us to carefully study the NGD method both theoretically and numerically. Theoretically, we start with a classical linear regression model. We find that both the learning rate and the network structure play significant roles in determining the NGD estimator's statistical efficiency. The resulting NGD estimator can be statistically as efficient as the global estimator, if the learning rate is sufficiently small and the network structure is weakly balanced, even if the data are distributed heterogeneously. Those interesting findings are then extended to general models and loss functions. Extensive numerical studies are presented to corroborate our theoretical findings. Classical deep learning models are also presented for illustration purpose.

#### Wasserstein Distance-Based Spectral Clustering with Application to Transaction Data

♦*Yingqiu Zhu<sup>1</sup>, Danyang Huang<sup>2</sup> and Bo Zhang<sup>2</sup>*

<sup>1</sup>University of International Business and Economics

<sup>2</sup>Renmin University of China  
inqzhu@uibe.edu.cn

With the rapid development of online payment platforms, it is now possible to record massive transaction data. The economic behaviors are embedded in the transaction data for merchants using these platforms. Therefore, clustering on transaction data significantly contributes to analyzing merchants' behavior patterns. This may help the platforms provide differentiated services or implement risk management strategies. However, traditional methods exploit transactions by generating low-dimensional features, leading to inevitable information loss. To fully use the transaction data, we utilize the empirical cumulative distribution of transaction amount to characterize merchants. Then, we adopt Wasserstein distance to measure the dissimilarity between any two merchants and propose the Wasserstein-distance-based spectral clustering (WSC) approach. Considering the sample imbalance in real datasets, we incorporate indices based on local outlier factors to improve the clustering performance. Furthermore, to ensure feasibility if the proposed method on large-scale datasets with limited computational resources, we propose a subsampling version of WSC. The associated theoretical properties are investigated to verify the efficiency of the proposed approach. The simulations and empirical study demonstrate that the proposed method outperforms feature-based methods in finding behavior patterns of merchants.

#### High-Dimensional Inference for Dynamic Treatment Effects

*Yuqian Zhang*  
yuqianzhang@ruc.edu.cn

Estimating dynamic treatment effects is a crucial endeavor in causal inference, particularly when confronted with high-dimensional confounders. Doubly robust (DR) approaches have emerged as promising tools for estimating treatment effects due to their flexibility. However, we showcase that the traditional DR approaches that only focus on the DR representation of the expected outcomes may fall short of delivering optimal results. In this paper, we propose a novel DR representation for intermediate conditional outcome models that leads to superior robustness guarantees. The proposed method achieves consistency even with high-dimensional confounders, as long as at least one nuisance function is appropriately parametrized for each exposure time and treatment path. Our results represent a significant step forward as they provide new robustness guarantees. The key to achieving these results is our new DR representation, which offers superior inferential performance while requiring weaker assumptions. Lastly, we confirm our findings in practice through simulations and a real data application.

## Session 23CHI142: Recent Advances in Spatio Temporal Modelling

### On Semiparametrically Dynamic Functional-Coefficient Autoregressive Spatio-Temporal Models with Irregular Location Wide Nonstationarity

Rongmao Zhang

Zhejiang University

rmzhang@zju.edu.cn

Nonlinear dynamic modeling of spatio-temporal data is often a challenge, especially due to irregularly observed locations and location-wide non-stationarity. In this talk, we introduce a semiparametric family of Dynamic Functional-coefficient Autoregressive Spatio-Temporal (DyFAST) models to address the difficulties. We specify the autoregressive smoothing coefficients depending dynamically on both a concerned regime and location so that the models can characterize not only the dynamic regime-switching nature but also the location-wide non-stationarity in real data. Different smoothing schemes are then proposed to model the dynamic neighboring-time interaction effects with irregular locations incorporated by (spatial) weight matrices. The first scheme popular in econometrics supposes that the weight matrix is pre-specified. We show that locally optimal bandwidths by a greedy idea popular in machine learning should be cautiously applied. Moreover, many weight matrices can be generated differently by data location features. Model selection is popular, but may suffer from loss of different candidate features. Our second scheme is thus to suggest a weight matrix fusion to let data combine or select the candidates with estimation done simultaneously. Simulation and empirical application to an EU energy market dataset are also presented to illustrate the usefulness of the proposed DyFAST models.

### One-Way or Two-Way Factor Model for Matrix Sequences?

Yong He<sup>1</sup>, ♦Xinbing Kong<sup>2</sup>, Lorenzo Trapani<sup>3</sup> and Long Yu<sup>4</sup>

<sup>1</sup>Shandong University

<sup>2</sup>Nanjing Audit University

<sup>3</sup>Nottingham University

<sup>4</sup>Shanghai University of Finance and Economics  
xinbingkong@126.com

In this talk, we investigate the issue of determining the dimensions of row and column factor spaces in matrix-valued data. Exploiting the eigen-gap in the spectrum of sample second moment matrices of the data, we propose a family of randomised tests to check whether a one-way or two-way factor structure exists or not. Our tests do not require any arbitrary thresholding on the eigenvalues, and can be applied with (virtually) no restrictions on the relative rate of divergence of the cross-sections to the sample sizes as they pass to infinity. Although tests are based on a randomization which does not vanish asymptotically, we propose a de-randomized, strong (based on the Law of the Iterated Logarithm) decision rule to choose in favour or against the presence of common factors. We use the proposed tests and decision rule in two ways. We further cast our individual tests in a sequential procedure whose output is an estimate of the number of common factors. Our tests are built on two variants of the sample second moment matrix of the data: one based on a row (or column) flattened version of the matrix-valued sequence, and one based on a projection-based method. Our simulations show that

### Multivariate Reduced-Rank Spatiotemporal Models

Dan Pu<sup>1</sup>, Kuangnan Fang<sup>1</sup>, ♦Wei Lan<sup>2</sup> and Qingzhao Zhang<sup>1</sup>

<sup>1</sup>Xiamen University

<sup>2</sup>Southwestern University of Finance and Economics  
lanwei@swufe.edu.cn

Multivariate spatiotemporal data arise frequently in practical applications, often involving complex dependencies across cross-sectional units, time points and multivariate variables. In the literature, few studies jointly model the dependence in three dimensions. To simultaneously model the cross-sectional, dynamic and cross-variable dependence, we propose a multivariate reduced-rank spatiotemporal model. By imposing the low-rank assumption on the cross-sectional influence matrix, the proposed model achieves substantial dimension reduction and has a nice interpretation, especially for financial data. Due to the innate endogeneity, we propose the quasi-maximum likelihood estimator (QMLE) to estimate the unknown parameters. A ridge-type ratio estimator is also developed to determine the rank of the cross-sectional influence matrix. We establish the asymptotic distribution of the QMLE and the rank selection consistency of the ridge-type ratio estimator. The proposed methodology is further illustrated via extensive simulation studies and two applications to a stock market dataset and an air pollution dataset.

### Design-Based Covariate Adjustment for Causal Inference with Interference and Noncompliance

Hanzhong Liu

Tsinghua University

lhz2016@tsinghua.edu.cn

In many social science and economic experiments, the experimental units may interact with each other such that the potential outcomes of one unit are affected by the treatment assignment of other units. Moreover, the experimental units may not comply with their treatment assignments, that is, the actual treatments received by the experimental units may be different from their treatments assigned. To address the interference issue, also known as dependent happenings, two-stage randomized experiments are widely used to estimate the direct and spillover effects. In this paper, we propose several covariate adjustment methods using regression or projection to efficiently estimate and infer the intention-to-treat and complier average direct and spillover effects in two-stage randomized experiments with both interference and noncompliance. We obtain the consistency and asymptotic normality of the covariate-adjusted intention-to-treat and complier average direct and spillover effects estimators and outline conditions under which they are asymptotically more efficient than the unadjusted Horvitz–Thompson and Hajek estimators. In addition, we propose nonparametric conservative covariance estimators to facilitate valid inferences. Our theory is design-based and does not require the regression model to be correctly specified. Numerical studies demonstrate the validity and efficiency gains of the proposed estimators.

## Session 23CHI145: Human-Centric Statistical Learning

### Privacy-Preserving Community Detection for Locally Distributed Multiple Networks

♦Xiao Guo<sup>1</sup>, Xiang Li<sup>2</sup>, Xiangyu Chang<sup>3</sup> and Shujie Ma<sup>4</sup>

<sup>1</sup>Northwest University, China

<sup>2</sup>Peking University, China

<sup>3</sup>Xi'an Jiaotong University, China

<sup>4</sup>University of California-Riverside, USA  
xiaoguo.stat@gmail.com

Modern multi-layer networks are commonly stored and analyzed in a local and distributed fashion because of the privacy, ownership,

## Abstracts

---

and communication costs. The literature on the model-based statistical methods for community detection based on these data is still limited. This paper proposes a new method for consensus community detection and estimation in a multi-layer stochastic block model using locally stored and computed network data with privacy protection. A novel algorithm named **privacy-preserving Distributed Spectral Clustering (ppDSC)** is developed. To preserve the edges' privacy, we adopt the *randomized response* (RR) mechanism to perturb the network edges, which satisfies the strong notion of *differential privacy*. The ppDSC algorithm is performed on the squared RR-perturbed adjacency matrices to prevent possible cancellation of communities among different layers. To remove the bias incurred by RR and the squared network matrices, we develop a two-step bias-adjustment procedure. Then we perform eigen-decomposition on the debiased matrices, aggregation of the local eigenvectors using an orthogonal Procrustes transformation, and  $k$ -means clustering. We provide theoretical analysis on the statistical errors of ppDSC in terms of eigen-vector estimation. In addition, the blessings and curses of network heterogeneity are well-explained by our bounds.

### Learning Multitask Gaussian Bayesian Networks

♦Shuai Liu<sup>1</sup>, Yixuan Qiu<sup>2</sup>, Baojuan Li<sup>3</sup>, Huaning Wang<sup>3</sup> and Xiangyu Chang<sup>1</sup>

<sup>1</sup>Xi'an Jiaotong University

<sup>2</sup>Shanghai University of Finance and Economics

<sup>3</sup>Air Force Medical University,  
hljliushuai@163.com

Bayesian network is a probabilistic graphical model representing the causal relationship between system variables. The typical problem is identifying multiple Bayesian networks for different sub-types of data related to a large-scale database. In this case, learning Bayesian networks in isolation can result in a precarious model since the observations of each sub-type may be relatively small. This article proposes a multitask Gaussian Bayesian network framework to learn multiple Bayesian networks from related small sample observations. We achieve this goal by treating each subject in a class of observations as a task and leveraging data from others. Since the subjects are in the same class, they share some similarities. Therefore, we introduce a shared prior covariance matrix among all tasks and utilize the hierarchical model structure to estimate multiple Gaussian Bayesian networks. For computing, we derive closed-form expressions for the complete likelihood function and its gradient and use the Monte Carlo expectation-maximization algorithm to search for the approximately best network structures efficiently. We assess the performance of the proposed method with extensive and systematic numerical experiments and highlight its usefulness in analyzing resting-state functional magnetic resonance imaging (rs-fMRI) data and facilitating the study of major depressive

### 2d-Shapley: a Framework for Fragmented Data Valuation

♦Zhihong Liu<sup>1</sup>, Hoang Anh Just<sup>2</sup>, Xiangyu Chang<sup>1</sup>, Xi Chen<sup>3</sup> and Ruoxi Jia<sup>2</sup>

<sup>1</sup>Xi'an Jiaotong University

<sup>2</sup>Virginia Tech

<sup>3</sup>New York University  
xiangyuchang@gmail.com

Data valuation—quantifying the contribution of individual data sources to certain predictive behaviors of a model—is of great importance to enhancing the transparency of machine learning and designing incentive systems for data sharing. Existing work has focused on evaluating data sources with the shared feature or sample space. How to evaluate fragmented data sources of which each only

contains partial features and samples remains an open question. We start by presenting a method to calculate the counterfactual of removing a fragment from the aggregated data matrix. Based on the counterfactual calculation, we further propose 2D-Shapley, a theoretical framework for fragmented data valuation that uniquely satisfies some appealing axioms in the fragmented data context. 2D-Shapley empowers a range of new use cases, such as selecting useful data fragments, providing interpretation for sample-wise data values, and fine-grained data issue diagnosis.

### 2d-Shapley: a Framework for Fragmented Data Valuation

♦Zhihong Liu<sup>1</sup>, Hoang Anh Just<sup>2</sup>, Xiangyu Chang<sup>1</sup> and Ruoxi Jia<sup>3</sup>

<sup>1</sup>Xi'an Jiaotong University

<sup>2</sup>Virginia Polytechnic Institute and State University

<sup>3</sup>Virginia Tech  
1251906353@qq.com

Data valuation—quantifying the contribution of individual data sources to certain predictive behaviors of a model—is of great importance to enhancing the transparency of machine learning and designing incentive systems for data sharing. Existing work has focused on evaluating data sources with the shared feature or sample space. How to evaluate fragmented data sources of which each only contains partial features and samples remains an open question. We start by presenting a method to calculate the counterfactual of removing a fragment from the aggregated data matrix. Based on the counterfactual calculation, we further propose 2D-Shapley, a theoretical framework for fragmented data valuation that uniquely satisfies some appealing axioms in the fragmented data context. 2D-Shapley empowers a range of new use cases, such as selecting useful data fragments, providing interpretation for sample-wise data values, and fine-grained data issue diagnosis.

### Toward a Fairness-Aware Scoring System for Algorithmic Decision-Making

♦Yi Yang<sup>1</sup>, Ying Wu<sup>2</sup>, Xiangyu Chang<sup>2</sup>, Mei Li<sup>3</sup> and Yong Tan<sup>4</sup>

<sup>1</sup>Xi'an Jiaotong-liverpool University

<sup>2</sup>Xi'an Jiaotong University

<sup>3</sup>University of Oklahoma

<sup>4</sup>University of Washington  
Yi.Yang@xjtu.edu.cn

Scoring systems, as a type of predictive model, have significant advantages in interpretability and transparency and facilitate quick decision-making. As such, scoring systems have been extensively used in a wide variety of industries such as healthcare. However, the fairness issues in these models have long been criticized, and the use of machine learning algorithms in the construction of scoring systems heightens this concern. In this paper, we propose a general framework to create fairness-aware, data-driven scoring systems. First, we develop a social welfare function that incorporates both efficiency and group fairness. Then, we transform the social welfare maximization problem into the risk minimization task in machine learning, and derive a fairness-aware scoring system with the help of mixed integer programming. Lastly, several theoretical bounds are derived for providing parameter selection suggestions. Our proposed framework provides a solution to address group fairness concerns in the development of scoring systems. It enables policymakers to set and customize their desired fairness requirements and other application-specific constraints. We test the proposed algorithm with several empirical data sets. Experimental evidence supports the effectiveness of the proposed scoring system in achieving the optimal welfare of stakeholders and in balancing the needs for interpretability, fairness, and efficiency.

## Session 23CHI147: Statistical Methods and Applications on Unstructured Data

### Trajectory Representation Learning with Multilevel Attention for Driver Identification

♦Mengyuan Li<sup>1</sup>, Yuanyuan Zhang<sup>2</sup>, Yaya Zhao<sup>1</sup>, Yalei Du<sup>2</sup> and Xiaoling Lu<sup>1</sup>

<sup>1</sup>Renmin University of China

<sup>2</sup>Beijing Baixingkefu Network Technology Co., Ltd.  
limengyuan@ruc.edu.cn

Massive trajectory data have originated from the development of positioning technology. Learning GPS trajectory representation to characterize a driver's driving style is a challenging task with important applications in many areas, including autonomous driving, auto insurance, advanced driver assistance systems, urban computing, and the internet of things. Few studies have considered the interactions between different factors. In this study, we propose a novel trajectory representation method based on a multilevel attention mechanism (ATTRaj2vec) and apply it to the task of driver identification. We use 1D CNN to summarize local features from sub-trajectories, including motion, spatial and temporal features. Then we utilize a multilevel attention mechanism to extract global features, aggregating the interactions of motion features with temporal and spatial features progressively. Additionally, we adopt multi-loss to optimize our model simultaneously, which consists of a softmax loss for driver classification and Siamese loss for making trajectories from the same driver more similar. Classification experimental results on a real-world automobile trajectory dataset demonstrated that our proposed model significantly outperforms existing baselines. Meanwhile, the proposed method provides significant gains in the trajectory clustering of unseen drivers.

### "this Crime is not that Crime"—Classification and Evaluation of Four Common Crimes

♦Ke Xu<sup>1</sup>, Hangyu Liu<sup>1</sup>, Fang Wang<sup>2</sup> and Hansheng Wang<sup>3</sup>

<sup>1</sup>University of International Business and Economics

<sup>2</sup>Shandong University

<sup>3</sup>Peking University  
xk@uibe.edu.cn

As the basis of criminal penalty, criminal conviction, integral to the protection of fundamental rights and freedom of people, constitutes the basis and the core issue of criminal trials. Based on the data published on China Judgments Online, we proposed 2 models to identify wrongly sentenced cases: a 2-stage model and 2 deep learning models. The 2-stage model used 3 keyword-extraction models and 5 classification models to extract important words and classify cases. The deep learning models employed 2 different neural network models. We tested both models on a validation data set of cases with changed verdicts and found that the 2-stage model was more effective, with important words often associated with abnormal cases. Both models were successful in identifying such cases, but the 2-stage model (aka. TF-IDF & ANN) outperformed the deep learning models. The study underscores the importance of criminal conviction in protecting fundamental rights and freedoms and highlights the potential of machine learning models in improving the accuracy of criminal trials.

### A Geometrical Model with Stochastic Error for Abnormal Motion Detection of Portal Crane Bucket Grab

♦Baichen Yu<sup>1</sup>, Xiao Wang<sup>2</sup> and Hansheng Wang<sup>3</sup>

<sup>1</sup>East China Normal University

<sup>2</sup>Qingdao University

<sup>3</sup>Peking University  
baichen.yu@stu.ecnu.edu.cn

Sea transportation is among the most important modes of transportation, accounting for more than 80% of international trades. Sea port infrastructure, especially portal cranes plays a crucial role among multiple components. Consequently, the safe and effective operation of portal cranes, including automatically monitoring the motions of a bucket grab, becomes a critically important issue. To potentially address this issue, we have developed a novel approach to estimate the swing angle of the portal crane using video images generated by a surveillance camera installed on the fly-jib head as the input. Next, a spatial geometric model with stochastic error is developed. The model describes the geometric relationship between the signals observed on the image plane and the actual bucket grab motion. A statistical model is used to describe the stochastic behavior of the bucket grab along with a novel iterative algorithm to estimate the unknown parameters. This enables us to estimate the swing angle in a timely manner and generate an alarm signal immediately. Numerical studies based on both simulated and real datasets are presented. Our method can be used to guarantee the day-to-day safe operations of portal cranes for transferring freight from the port to cargo ships, and vice versa.

### Road Network Enhanced Human Activity Recognition with Spatial-Temporal Transformer

Yaya Zhao

Renmin University of China  
zhaoyaya@ruc.edu.cn

This study aims to tackle the SHL recognition challenge, which involves the recognition of location-independent and user-independent activities using smartphone sensors. To effectively address the long-range temporal aspect of this problem, we propose a novel approach called the road network enhanced transformer-based model. Our model is designed to classify eight different activities accurately. In this study, we begin by conducting data pre-processing to extract a set of relevant and discriminative features. Furthermore, we incorporate road network information into our transformer-based model. By leveraging the road network data, our model gains a better understanding of the spatial context in which the activities occur. Through extensive experimentation and evaluation, our proposed method achieves a promising accuracy score of 0.80 on the validation dataset.

## Session 23CHI148: New Developments on Design of Experiments and Sampling Methods

### Adaptive Order-of-Addition Experiments via the Quick-Sort Algorithm

Dennis K. J. Lin<sup>1</sup> and ♦Jianbin Chen<sup>2</sup>

<sup>1</sup>Purdue University

<sup>2</sup>Beijing Institute of Technology  
chenjianbinlu@163.com

The order-of-addition (OofA) experiment has received a great deal of attention in the recent literature. The primary goal of the OofA experiment is to identify the optimal order in a sequence of m components. All the existing methods are model-dependent and are limited to a small number of components. The appropriateness of the resulting optimal order heavily depends on (a) the correctness of the underlying assumed model, and (b) the goodness of model fitting. Moreover, these methods are not applicable to dealing with large m (e.g., m7). With this in mind, this article proposes an efficient

## Abstracts

---

adaptive methodology, building upon the quick-sort algorithm, to explore the optimal order without any model specification. Compared to the existing work, the run sizes of the proposed method needed to achieve the optimal order are much smaller. Theoretical supports are given to illustrate the effectiveness of the proposed method. The proposed method is able to obtain the optimal order for large  $m$  (e.g.,  $m=20$ ). Numerical experiments are used to demonstrate the effectiveness of the proposed method.

### Interaction Effects in Pairwise Ordering Model

♦*Chunyan Wang<sup>1</sup> and Dennis Lin<sup>2</sup>*

<sup>1</sup>aCenter for Applied Statistics and School of Statistics, Renmin University of China

<sup>2</sup>bDepartment of Statistics, Purdue University, West Lafayette  
chunyanwang@ruc.edu.cn

In an order-of-addition (OoA) experiment, the response is a function of the addition order of components. The key objective of the OoA experiments is to find the optimal order of addition. The most popularly used model for OoA experiments is perhaps the pairwise ordering (PWO) model, which assumes that the response can be fully accounted by the pairwise ordering of components. Recently, Mee (2020) extended the PWO model by adding the interactions of PWO factors, to account for variations caused by the ordering of sets of three or more components, where the interaction term is defined by the multiplication of two PWO factors. This paper introduces a novel class of conditional PWO effect to study the interaction effect between PWO factors. The advantages of the proposed interaction terms are studied. Based on these conditional effects, a new model is proposed. The optimal order of addition can be straightforwardly obtained via the proposed model.

### Subsampling Markov Chain Monte Carlo Algorithm Based on Energy Distance

♦*Sumin Wang<sup>1</sup>, Fasheng Sun<sup>2</sup> and Min-Qian Liu<sup>1</sup>*

<sup>1</sup>Nankai University

<sup>2</sup>Northeast Normal University  
wangsm088@nankai.edu.cn

Markov chain Monte Carlo (MCMC) algorithms are generally regarded as the gold standard technique for Bayesian inference. They are theoretically well-understood and conceptually simple to apply in practice. The drawback of MCMC is that performing exact inference generally requires all of the data to be processed at each iteration of the algorithm. For large datasets, the computational cost of MCMC can be prohibitive, which has led to recent developments in scalable Monte Carlo algorithms that have a significantly lower computational cost than standard MCMC. In this article, we focus on a particular class of scalable Monte Carlo algorithms, which utilizes data subsampling techniques to reduce the per-iteration cost of MCMC. The subsampling process is guided by the fidelity to the observed data, as measured by the energy distance of two sets of samples. The resulting algorithm is a generic and flexible approach that preserves the simplicity of the Metropolis-Hastings algorithm. Even though exactness is lost, i.e. the chain distribution approximates the posterior, we study and quantify theoretically this bias and show on a diverse set of examples that it yields excellent performances when the computational budget is limited.

### Construction of Orthogonal Maximin Distance Designs

♦*Wenlong Li<sup>1</sup>, Yubin Tian<sup>1</sup> and Min-Qian Liu<sup>2</sup>*

<sup>1</sup>Beijing Institute of Technology

<sup>2</sup>Nankai University  
wenlongli@mail.nankai.edu.cn

Maximin distance designs and orthogonal designs are two attractive classes of space-filling designs for computer experiments, but their theoretical constructions are challenging, especially the construction of optimal designs in terms of both the maximin distance and orthogonality criteria. This paper presents a systematic method for constructing orthogonal maximin distance designs with flexible numbers of runs and factors. The method is carried out by rotating the subarrays of a saturated two-level regular design in Yates order or its circular shifting version. The principal objective is to construct high-level designs from two-level designs, and the method is effective because the performance of high-level designs is determined by that of two-level designs under both the maximin distance and orthogonality criteria. The proposed method is also generalized by rotating the subarrays of a saturated two-level nonregular design such that the resulting designs have flexible run sizes. Comparison results reveal that the resulting orthogonal designs are well worthy of recommendation under the maximin distance criterion. An illustrative example is provided to show that the proposed designs have a good two-dimensional stratification property.

### Session 23CHI162: New Methods and Applications of Reinforcement Learning for Complex Data

#### Value Enhancement of Reinforcement Learning via Efficient and Robust Trust Region Optimization

Chengchun Shi<sup>1</sup>, Zhenglin Qi<sup>1</sup>, ♦Jianing Wang<sup>2</sup> and Fan Zhou<sup>1</sup>

<sup>1</sup>Advisor

<sup>2</sup>me

jianing.wang@163.sufe.edu.cn

Motivated by high stake domains such as mobile health studies with limited and pre-collected data, in this paper, we study offline reinforcement learning methods. To efficiently use these datasets for policy optimization, we propose a novel value enhancement method to improve the performance of a given initial policy computed by existing state-of-the-art RL algorithms. Specifically, when the initial policy is not consistent, our method will output a policy whose value is no worse and often better than that of the initial policy. When the initial policy is consistent, under some mild conditions, our method will yield a policy whose value converges to the optimal one at a faster rate than the initial policy, achieving the desired “value enhancement” property. The proposed method is generally applicable to any parametrized policy that belongs to certain pre-specified function class (e.g., deep neural networks). Extensive numerical studies are conducted to demonstrate the superior performance of our method.

#### Doubly Inhomogeneous Reinforcement Learning

♦Liyuan Hu<sup>1</sup>, Mengbing Li<sup>2</sup>, Chengchun Shi<sup>1</sup>, Zhenke Wu<sup>2</sup> and Piotr Fryzlewicz<sup>1</sup>

<sup>1</sup>London School of Economics and Political Science

<sup>2</sup>University of Michigan, Ann Arbor

l.hu11@lse.ac.uk

This paper studies reinforcement learning in doubly inhomogeneous environments under temporal non-stationarity and subject heterogeneity. In many applications, it is commonplace to encounter datasets generated by system dynamics that change over time and population, challenging high-quality sequential decision making. In this paper, we propose an original algorithm to determine the “best data rectangles” that display similar dynamics over time and population to speed up the policy learning, which alternates between change point detection and cluster identification. Our method is

general, multiply robust and efficient. Empirically, we demonstrate the usefulness of our method through extensive numerical experiments.

### Damped Anderson Mixing for Deep Reinforcement Learning: Acceleration, Convergence, and Stabilization

*Ke Sun<sup>1</sup>, Yafei Wang<sup>1</sup>, Yi Liu<sup>1</sup>, Bo Pan<sup>1</sup>, Yingnan Zhao<sup>2</sup>, Shangling Jui<sup>3</sup>, Bei Jiang<sup>1</sup> and ♦Linglong Kong<sup>1</sup>*

<sup>1</sup>University of Alberta

<sup>2</sup>Harbin Institute of Technology

<sup>3</sup>Huawei Technologies Ltd

lkong@ualberta.ca

Anderson mixing has been heuristically applied to reinforcement learning (RL) algorithms for accelerating convergence and improving the sampling efficiency of deep RL. In this paper, we provide deeper insights into a class of acceleration schemes built on Anderson mixing that improve the convergence of deep RL algorithms. Our main results establish a connection between Anderson mixing and quasi-Newton methods and prove that Anderson mixing increases the convergence radius of policy iteration schemes by an extra contraction factor. The key focus of the analysis roots in the fixed-point iteration nature of RL. We further propose a stabilization strategy by introducing a stable regularization term in Anderson mixing and a differentiable, non-expansive MellowMax operator that can allow both faster convergence and more stable behavior. Extensive experiments demonstrate that our proposed method enhances the convergence, stability, and performance of RL algorithms.

### Application of Distributional Reinforcement Learning in Ride-Sharing Industry

*Fan Zhou*

zhoufan@mail.shufe.edu.cn

We develop a multi-objective distributional reinforcement learning framework for improving order dispatching on large-scale ride-hailing platforms. Compared with traditional RL-based approaches that focus on drivers' income, the proposed framework also accounts for the spatiotemporal difference between the supply and demand networks. Specifically, we model the dispatching problem as a two-objective Semi-Markov Decision Process (SMDP) and estimate the relative importance of the two objectives under some unknown existing policy via Inverse Reinforcement Learning (IRL). Then, we combine Implicit Quantile Networks (IQN) with the traditional Deep Q-Networks (DQN) to jointly learn the two return distributions and adjusting their weights to refine the old policy through on-line planning and achieve a higher supply-demand coherence of the platform. We conduct large-scale dispatching experiments to demonstrate the remarkable improvement of proposed approach on the platform's efficiency.

### Session 23CHI164: Complex Statistical Modelling and Testing

#### Test of the Latent Dimension of a Spatial Blind Source Separation Model

*Christoph Muehlmann, François Bachoc, Klaus Nordhausen and ♦Mengxi Yi  
mxyi@bnu.edu.cn*

We assume a spatial blind source separation model in which the observed multivariate spatial data is a linear mixture of latent spatially uncorrelated random fields containing a number of pure white noise

components. We propose a test on the number of white noise components and obtain the asymptotic distribution of its statistic for a general domain. We also demonstrate how computations can be facilitated in the case of gridded observation locations. Based on this test, we obtain a consistent estimator of the true dimension. Simulation studies and an environmental application in the Supplemental Material demonstrate that our test is at least comparable to and often outperforms bootstrap-based techniques, which are also introduced in this paper.

#### Testing Sufficiency for Transfer Learning

*♦Ziqian Lin<sup>1</sup>, Yuan Gao, Feifei Wang<sup>2</sup> and Hansheng Wang*

<sup>1</sup>linziqian@stu.pku.edu.cn

<sup>2</sup>feifei.wang@ruc.edu.cn

linziqian@stu.pku.edu.cn

Modern statistical analysis often encounters high dimensional models but with limited sample sizes. This makes the target data based statistical estimation very difficult. Then how to borrow information from another large-sized source data for more accurate target model estimation becomes an interesting problem. This leads to the useful idea of transfer learning. Various estimation methods have been developed recently. In this work, we study transfer learning from a different perspective. Specifically, we consider here the problem of testing for transfer learning sufficiency. By transfer learning sufficiency (denoted as the null hypothesis), we mean that, with the help of the source data, the useful information contained in the feature vectors of the target data can be sufficiently extracted for predicting the interested target response. Therefore, the rejection of the null hypothesis implies that information useful for prediction remains in the feature vectors of the target data and thus calls for further exploration. To this end, we develop a novel testing procedure and a centralized and standardized test statistic, whose asymptotic null distribution is analytically derived. Simulation studies are presented to demonstrate the finite sample performance of the proposed method. A deep learning related real data example is presented for illustration purpose.

#### Consistent Selection of the Number of Groups in Panel Models via Sample-Splitting

*♦Zhe Li<sup>1</sup>, Xuening Zhu and Changliang Zou*

<sup>1</sup>Fudan University

zheli20@fudan.edu.cn

Group number selection is a key question for group panel data modelling. In this work, we develop a cross validation method to tackle this problem. Specifically, we split the panel data into a training dataset and a testing dataset on the time span. We first use the training dataset to estimate the parameters and group memberships. Then we apply the fitted model to the testing dataset and then the group number is estimated by minimizing certain loss function values on the testing dataset. We design the loss functions for panel data models either with or without fixed effects. The proposed method has two advantages. First, the method is totally data-driven thus no further tuning parameters are involved. Second, the method can be flexibly applied to a wide range of panel data models. Theoretically, we establish the estimation consistency by taking advantage of the optimization property of the estimation algorithm. Experiments on a variety of synthetic and empirical datasets are carried out to further illustrate the advantages of the proposed method.

#### Distributed Estimation and Inference for Spatial Autoregression Model with Large Scale Networks

*♦Yimeng Ren, Zhe Li, Yuan Gao and Hansheng Wang*

ymren22@m.fudan.edu.cn

## Abstracts

The rapid growth of online network platforms generates large-scale network data and it poses great challenges for statistical analysis using the spatial autoregression (SAR) model. In this work, we develop a novel distributed estimation and statistical inference framework for the SAR model on a distributed system. We first propose a distributed network least squares approximation (DNLSA) method. This enables us to obtain a one-step estimator by taking a weighted average of local estimators on each worker. Afterwards, a refined two-step estimation is designed to further reduce the estimation bias. For statistical inference, we utilize a random projection method to reduce the expensive communication cost. Theoretically, we show the consistency and asymptotic normality of both the one-step and two-step estimators. In addition, we provide theoretical guarantee of the distributed statistical inference procedure. The theoretical findings and computational advantages are validated by several numerical simulations implemented on the Spark system. Lastly, an experiment on the Yelp dataset further illustrates the usefulness of the proposed methodology.

### Session 23CHI168: Recent Advances in Sequencing Data Analysis, Reinforcement Learning and Missing Data Handling

#### Statistical Methods for Allele-Specific Expression Analysis using Single-Cell Rna-Seq Data

Rui Xiao

University of Pennsylvania  
rxiao@pennmedicine.upenn.edu

Abstract: Allele-specific gene expression (ASE) analysis, an alternative and complementary approach to eQTL analysis, is a powerful tool for identifying variation in gene expression. ASE quantifies the relative expression of two alleles in a diploid individual, and the imbalance of expression of the two alleles may explain phenotypic variation and disease pathophysiology. ASE is driven by cis-regulatory variants located near a gene. Since the two alleles used to measure ASE are expressed in the same cellular environment and genetic background, they can serve as internal controls and eliminate the influence of trans-acting genetic and environmental factors. In this talk, I will focus on statistical methods for ASE analysis using RNA sequencing (RNA-seq) and single-cell RNA-seq (scRNA-seq) data that we recently developed. Specifically, I will first introduce a statistical model for detection of gene-level ASE across multiple individuals in a population under one clinical condition, as well as ASE difference between two clinical conditions. ASE patterns may vary across cell types. To better identify cellular targets of disease, we will next introduce a recently developed statistical method to characterize cell-type-specific ASE in bulk RNA-seq data by incorporating cell type composition information inferred from external scRNA-seq data. This method is extended to

#### Learning to Make Adherence-Aware Recommendations

◆Guanting Chen<sup>1</sup>, Xiaocheng Li<sup>2</sup>, Chunlin Sun<sup>3</sup> and Hanzhao Wang<sup>2</sup>

<sup>1</sup>University of North Carolina at Chapel Hill

<sup>2</sup>Imperial College Business School

<sup>3</sup>Stanford University  
guanting@unc.edu

As AI systems continue to make recommendations for human decision-making, it is frequently observed that human agents sometimes disregard these recommendations. In such cases, it may be beneficial for the AI system to refrain from providing the

optimal recommendation, which assumes perfect adherence from the agent. We propose a decision-making model that considers adherence-aware recommendations, accounting for the varying levels of adherence exhibited by human agents across different states and actions. Aside from the model, we also introduce accountable and near-optimal reinforcement learning algorithms specifically designed to address adherence-aware recommendations.

#### Rna Sequencing Differential Analysis using Deep-Learning Algorithms

Shijian Deng, Zedian Xie and ♦Dongmei Li

University of Rochester  
Dongmei\_Li@urmc.rochester.edu

RNA sequencing has been widely used in biomedical research to identify novel genes and cell types associated with different treatment conditions or diseases. With the decreasing of sequencing costs in recent years, large amount of sequencing data has been generated from NIH funded consortium grants across multiple institutions. Given the availability of large sample size in sequencing data, application of machine learning and deep learning algorithms into RNA sequencing data analyses getting more popular in recent years. Our study focuses on examining the performance of machine learning and deep learning algorithms in RNA sequencing differential analysis comparing with popular statistical methods, through both simulation studies and real data examples. The performance of different algorithms will be indicated by the false discovery rate control, power, and stability.

#### Leveraging Real World Evidence in Regulatory Submissions and Handling Missing Outcome in Real World Data

♦Jingyuan Yang, Terry Boodhoo and Hongyan Qiao

AbbVie  
jingyuan.yang@abbvie.com

For sponsors of medical products, providing data from prospective randomized controlled trials has long been considered the gold standard to demonstrate the safety and efficacy of a regulated product. However, for many medical devices, practical limitations related to the device or disease condition require alternative approaches to prospective randomized controlled trials and increased flexibility in trial design and statistical analysis. Advances in the availability of real-world data (RWD) sources, such as electronic health records, registries, medical claims, pharmacy data and feedback from wearables and mobile technology, have increased the potential to generate robust real-world evidence (RWE). RWE is clinical evidence regarding the usage, and benefits and risks, of a medical product derived from the analysis of RWD, which can be leveraged to support regulatory decisions. Overcoming bias due to non-randomized study design and handling missing data are challenging issues when analyzing RWD. This presentation will cover a statistical method we propose to utilize propensity score stratification and multiple imputation jointly to address these challenges in RWD analysis. 对于医疗产品公司，提供来自前瞻性随机对照试验的数据一直被认为证明受监管产品安全性和有效性的黄金标准。然而对于许多医疗器械，与设备或疾病状况相关的实际因素制约了前瞻性随机对照试验的可行性，所以需要替代方法来提高试验设计和统计分析的灵活性。真实世界数据 (RWD) 可用性的进步，包括但不限于电子健康记录、医学数据库、医疗索赔、药房数据以及来自可穿戴设备和移动技术的反馈等，增加了生成可靠真实世界证据 (RWE) 的潜力。RWE 是从 RWD 分析中得出的医疗产品的使用、益处和风险的临床证据，可以用来支持监管决策。在分析 RWD 时，克服非随机研究设计和处理缺失数据导致的偏差是具有挑战性的问题。倾向性评分 (PS) 通常用于解

决偏差问题，多重插补是解决缺失数据问题的广泛应用方法。然而，大多数现有文献侧重于处理 PS 模型协变量中的缺失数据，而不是 RWD 中结果变量中的缺失数据。排除缺少结果变量的受试者的完整分析将使分析集偏离意向治疗人群，并可能使分析结果产生偏差。当同时需要 PS 建模和处理结果变量中的缺失数据时，找到最佳的处理方式将是一个新的挑战。本演讲将介绍我们建议的联用 PS 分层和结果变量缺失数据多重插补的统计方法，以应对 RWD 分析中的这一挑战。

## Session 23CHI3: Computational Approaches to Single-Cell Genomics and Spatial-Omics Data Analysis and Clinical Applications

### Large-Scale Single-Cell-Based Deconvolution of Pan-Liver Diseases and Spatial Transcriptomics Identified a Novel Cell-Based Marker in Liver Cancer

Bin Chen

Michigan State University  
chenbi12@msu.edu

Hepatocellular carcinoma (HCC) is the third leading cause of cancer-related deaths worldwide due to its late diagnosis and limited effective treatment options. Morbidity is highest in HCC patients with underlying chronic liver diseases (CLDs). The majority of HCC develops from chronic liver fibrosis and/or cirrhosis, which predominantly results from chronic liver inflammation. Cellular heterogeneity in disease progression from CLDs to HCC remains poorly understood. Various genomics, proteomics, metabolomics, and integrative omics approaches have been investigated in attempts to identify prognostic and diagnostic markers for HCC. These efforts provide expanded options to established serum diagnostic markers such as alpha-fetoprotein (AFP) and glypican-3 (GPC3). However, none of these studies have focused on cell type specific markers. While large-scale scRNA-seq profiling of tissues across different stages of liver diseases is currently limited, the combination of bulk RNA-seq data with rich clinical annotation, established gene markers of reference cell types, and advanced computational methods enabled our integrative approach to understanding how cell-type composition changes during liver disease progression. I will talk about the discovery of a novel, non-invasive and actionable diagnostic marker of HCC through large-scale single-cell-based deconvolution of bulk tissues, followed by extensive preclinical and clinical validations. I will also share the usage

### Fastmix: Making Cell Type-Specific Biomarker Inference with Flowcytometry Data and Gene Expressions without Deconvolution

Yun Zhang<sup>1</sup>, Hao Sun<sup>2</sup>, Aishwarya Mandava<sup>3</sup>, Brian Aebermann<sup>3</sup>, Tobias Kollmann<sup>4</sup>, Richard Scheuermann<sup>3</sup>, Xing Qiu<sup>2</sup> and Yu Qian<sup>5</sup>

<sup>1</sup>J. Craig Venter Institute

<sup>2</sup>University of Rochester

<sup>3</sup>J. Craig Venter Institute

<sup>4</sup>University of Western Australia

<sup>5</sup>J. Craig Venter Institute,  
xing\_qiu@urmc.rochester.edu

We developed a novel multiomics analytics pipeline, FastMix, which integrates flow cytometry, bulk transcriptomics, and clinical covariates for identifying cell type-specific gene expression signatures and biomarker genes. FastMix addresses the “large p, small n” problem in multiomics data analysis via a linear mixed-effects model (LMER) for both cross-sectional and longitudinal studies.

Its novel moment-based estimator not only reduces bias in parameter estimation but also is more efficient than iterative optimization. Simulation studies showed that FastMix produced smaller type I/II errors than competing methods. Validation using real data of two vaccine studies showed that FastMix identified a consistent set of signature genes as in independent single cell RNA-seq analysis, producing additional interesting findings.

### Asgard is a Single-Cell Guided Pipeline to Aid Repurposing of Drugs

Bing He, Yao Xiao, Haodong Liang, Qianhui Huang, Yuheng Du, Yijun Li, David Garmire, Duxin Sun and <sup>◆</sup>Lana Garmire  
lgarmire@umich.edu

Single-cell RNA sequencing technology has enabled in-depth analysis of intercellular heterogeneity in various diseases. However, its full potential for precision medicine has yet to be reached. Towards this, we propose A Single-cell Guided Pipeline to Aid Repurposing of Drugs (ASGARD) that defines a drug score to recommend drugs by considering all cell clusters to address the intercellular heterogeneity within each patient. ASGARD shows significantly better average accuracy on single-drug therapy compared to two bulk-cell-based drug repurposing methods. We also demonstrate that it performs considerably better than other cell cluster-level predicting methods. In addition, we validate ASGARD using the drug response prediction method TRANSACT with Triple-Negative-Breast-Cancer patient samples. We find that many top-ranked drugs are either approved by the Food and Drug Administration or in clinical trials treating corresponding diseases. In conclusion, ASGARD is a promising drug repurposing recommendation tool guided by single-cell RNA-seq for personalized medicine. ASGARD is free for educational use at <https://github.com/lanagarmire/ASGARD>.

### Dissection of the Cell Interaction Landscapes Based on Single-Cell Spatial Transcriptome Data with Artificial Intelligence

Runze Li and <sup>◆</sup>Xuerui Yang

Tsinghua University  
yangxuerui@tsinghua.edu.cn

Recent advances in spatial transcriptomics have made it possible to simultaneously profile the intrinsic gene expression levels and the spatial organizations of hundreds to thousands of cells in complex tissues. It remains a major challenge to fully digest the single-cell spatial transcriptome profiles and recover the complete cell interaction networks simply from these incomplete, noisy, sparse snapshots of cellular organizations. We introduce a deep-learning strategy, named DeepLinc, for mining the latent features of single-cell spatial transcriptome profiles and eventually rebuilding comprehensive cell interaction networks. Specifically, DeepLinc combines variational graph autoencoders (VGAEs) and generative adversarial networks (GANs) to decode data, learn hidden features, and reconstruct cell interaction networks. When applied to 4 datasets of spatial transcriptome profiles generated by 3 different technologies, DeepLinc demonstrated its high efficiency in learning from such imperfect and incomplete data, filtering false interactions, and imputing missing distal and local interactions. The reconstructed full networks of cell interactions exhibited high physiological relevance. Key genes that have potentially contributed to the cell interaction landscapes were inferred by interrogating the DeepLinc pipeline for features with heavy weights. Finally, the latent representations learned by DeepLinc unveil spatially coded cell heterogeneity, which reflects physiologically relevant multi-cellular domains.

## Abstracts

---

### Session 23CHI43: Recent Advancements in Bayesian Methods for Causal Inference

#### A Bayesian Non-Parametric Approach for Causal Mediation with a Post-Treatment Confounder

♦Woojung Bae, Michael Daniels and Michael Perri

University of Florida

matt.woojung@gmail.com

We propose a new Bayesian non-parametric (BNP) method for estimating the causal effects of mediation in the presence of a post-treatment confounder. We specify an enriched Dirichlet process mixture (EDPM) to model the joint distribution of the observed data (outcome, mediator, post-treatment confounder, treatment, and baseline confounders). For identifiability, we use the extended version of the standard sequential ignorability as introduced in hong2022posttreatment. The observed data model and causal identification assumptions enable us to estimate and identify the causal effects of mediation, *i.e.*, the natural direct effects (NDE), and indirect effects (NIE). Our method enables easy computation of NDE and NIE for a subset of confounding variables and addresses missing data through data augmentation under the assumption of ignorable missingness. We conduct simulation studies to assess the performance of our proposed method. Furthermore, we apply this approach to evaluate the causal mediation effect in the Rural LITE trial, demonstrating its practical utility in real-world scenarios.

#### In Nonparametric and High-Dimensional Models, Bayesian Ignorability is an Informative Prior

Antonio Linero

University of Texas at Austin

antonio.linero@austin.utexas.edu

In problems with large amounts of missing data one must model two distinct data generating processes: the outcome process, which generates the response, and the missing data mechanism, which determines the data we observe. Under the ignorability condition of Rubin (1976), however, likelihood-based inference for the outcome process does not depend on the missing data mechanism so that only the former needs to be estimated; partially because of this simplification, ignorability is often used as a baseline assumption. We study the implications of Bayesian ignorability in the presence of high-dimensional nuisance parameters and argue that ignorability is typically incompatible with sensible prior beliefs about the amount of confounding bias. We show that, for many problems, ignorability directly implies that the prior on the selection bias is tightly concentrated around zero. This is demonstrated on several models of practical interest, and the effect of ignorability on the posterior distribution is characterized for high-dimensional linear models with a ridge regression prior. We then show both how to build high-dimensional models that encode sensible beliefs about the confounding bias and also show that under certain narrow circumstances ignorability is less problematic.

#### Bayesian Semiparametric Models for Dynamic Treatment Strategies with Incomplete Covariate Information

Arman Organisian

Brown University

arman\_organisian@brown.edu

We develop a Bayesian semiparametric model for assessing the impact of dynamic treatment rules (DTRs) on survival in the presence of missing time-varying covariates. Our work is motivated by a study of patients diagnosed with pediatric acute myeloid leukemia (AML). The data are from a phase III clinical trial in which patients move through a sequence of four treatment courses. At each course,

a decision is made to administer anthracyclines (ACT). Since ACT is cardiotoxic, left ventricular ejection fraction (EF) is sometimes - but not always - measured and used to help inform the ACT decision at each course. The inconsistent, and likely non-random, monitoring induces informative missingness in EF. In addition to missingness, patients may die or be withdrawn from the study before ever completing the sequence - leaving us with incomplete survival time information. We frame the problem in terms of a joint DTR that outputs both an EF monitoring decision and, conditional on the resulting information set, an ACT treatment decision. Gamma Process priors are used to flexibly model transitions between treatment courses and death under hypothetical DTRs in continuous time. A g-computation procedure is used to compute posterior marginal survival probabilities under hypothetical DTRs.

#### A Bayesian Machine Learning Approach for Estimating Heterogeneous Survivor Causal Effects: Applications to a Critical Care Trial

Xinyuan Chen<sup>1</sup>, Michael Harhay<sup>2</sup>, ♦Guangyu Tong<sup>3</sup> and Fan Li<sup>3</sup>

<sup>1</sup>Mississippi State University

<sup>2</sup>University of Pennsylvania

<sup>3</sup>Yale University

guangyu.tong@yale.edu

Assessing heterogeneity in the effects of treatments has become increasingly popular in the field of causal inference and carries important implications for clinical decision-making. While extensive literature exists for studying treatment effect heterogeneity when outcomes are fully observed, there has been limited development in tools for estimating heterogeneous causal effects when patient-centered outcomes are truncated by a terminal event, such as death. Due to mortality occurring during study follow-up, the outcomes of interest are unobservable, undefined, or not fully observed for many participants, in which case principal stratification is an appealing framework to draw valid causal conclusions. Motivated by the Acute Respiratory Distress Syndrome Network (ARDSNetwork) ARDS respiratory management (ARMA) trial, we developed a flexible Bayesian machine learning approach to estimate the average causal effect and heterogeneous causal effects among the always-survivors stratum when clinical outcomes are subject to truncation. We adopted Bayesian additive regression trees (BART) to flexibly specify separate mean models for the potential outcomes and latent stratum membership. In the analysis of the ARMA trial, we found that the low tidal volume treatment had an overall benefit for participants sustaining acute lung injuries on the outcome of time to returning home, but substantial heterogeneity in treatment effects.

### Session 23CHI89: Modern Learning Methods for Causal Inference and Survey Inference

#### Inference for Treatment Effects on Multiple Derived Outcomes

♦Yumou Qiu<sup>1</sup>, Jiarui Sun<sup>2</sup> and Xiao-Hua Zhou<sup>2</sup>

<sup>1</sup>Iowa State University

<sup>2</sup>Peking University

yumouqiu@iastate.edu

In many applications, the interest is in treatment effects on random quantities of subjects, where those random quantities are not directly observable but can be estimated based on data from each subject. In this paper, we propose a general framework for studying causal inference of this type of problems under a hierarchical data generation setting. The identifiability of causal parameters of interest is shown under a condition on the biasedness of subject level

estimates and an ignorability condition on the treatment assignment. Estimation of the treatment effects is constructed by inverse propensity score weighting on the estimated subject level parameters. A multiple testing procedure able to control the false discovery proportion is proposed to identify the nonzero treatment effects. Theoretical results are developed to investigate the proposed procedure, and numerical simulations are carried out to evaluate its empirical performance. A case study of medication effects on brain functional connectivity of patients with Autism spectrum disorder (ASD) using fMRI data is conducted to demonstrate the utility of the proposed method.

### **Probability Weighted Clustered Coefficients Regression Models in Complex Survey Sampling**

♦*Mingjun Gang<sup>1</sup>, Xin Wang<sup>2</sup>, Zhonglei Wang<sup>1</sup> and Wei Zhong<sup>1</sup>*

<sup>1</sup>Xiamen University

<sup>2</sup>San Diego State University

gangmingjun@126.com

Regression models are studied in survey data and are widely used to construct model-based estimators. Oftentimes, the relationships vary across different subjects or domains. Identifying a correct model structure with consideration of sampling weights is essential in making inferences and estimating population parameters. In this work, we propose the weighted clustered coefficients regression models for grouping covariate effects for survey data. The new method uses a weighted loss function and pairwise penalties on all pairs of observations. An algorithm based on the alternating direction method of multiplier algorithm is developed to obtain the estimates. We also study the theoretical properties of the estimator under the survey sampling setup. In the simulation study, the empirical performance of the proposed estimator is compared to the method without sampling weights, which suggests that sampling weights are important in identifying clusters in regression models.

### **Weighted Euclidean Balancing for a Matrix Exposure in Estimating Causal Effect**

*Juan Chen and ♦Yingchun Zhou*

East China Normal University

yczhou@stat.ecnu.edu.cn

In many scientific fields such as biology, psychology and sociology, there is an increasing interest in estimating the causal effect of a matrix exposure on an outcome. Although exact balancing and approximate balancing methods have been proposed for multiple balancing constraints, due to the large number of constraints when the treatment is a matrix, it is difficult to achieve exact balance or to select the threshold parameters for approximate balancing methods. To meet this challenge, the weighted Euclidean balancing method is proposed, which approximately balances covariates from an overall perspective. Both parametric and nonparametric methods are proposed to estimate the causal effect of a matrix treatment and theoretical properties of the two estimations are provided. Furthermore, the simulation results show that the proposed method outperforms other methods in various cases. Finally, the method is applied to investigating the causal relationship between children's participation in various training courses and their IQ. The results show that the weekly course hour of attending hands-on practice courses for children 6-9 years old has a positive impact on children's IQ with statistical significance.

### **Inferring Causal Effects on Survival Probability in a Double-Conounded Setting**

♦*Yang Bai<sup>1</sup> and Yifan Cui<sup>2</sup>*

<sup>1</sup>National University of Singapore

<sup>2</sup>Zhejiang University  
yang\_bai@u.nus.edu

In observational studies with survival data, researchers are often interested in studying the causal effect of exposure on survival probability. In addition to the unmeasured confounding, dependent censoring might pose additional challenges. We develop various bounds for causal effects on survival probability in this double-confounded setting. We provide a discussion on these bounds under different censoring contexts. Extensive simulations are also performed.

### **Session 23CHI9: Some Recent Developments in Deep Learning**

#### **Generative Conformal Prediction**

♦*Cheng Li<sup>1</sup>, Guohao Shen<sup>2</sup>, Yuanyuan Lin<sup>3</sup> and Jian Huang<sup>4</sup>*

<sup>1</sup>1155151973@link.cuhk.edu.hk

<sup>2</sup>guohao.shen@polyu.edu.hk

<sup>3</sup>ylin@cuhk.edu.hk

<sup>4</sup>j.huang@polyu.edu.hk

lic318320@gmail.com

We present a novel method for building conformal prediction intervals using a deep generative model. Unlike conformalized quantile regression, which approximates the conditional quantile function, our method learns a conditional generator function from the training data. We then use the generator to produce samples and calculate conformity scores on a calibration set, which allow us to construct conformal prediction intervals for new inputs. Our method can adapt to data heteroscedasticity. It also offers a natural way to incorporate unlabelled data in a semi-supervised setting. We prove the validity of our method under both supervised and semi-supervised settings. We demonstrate its superior performance over existing methods through simulations and real data examples, and show the benefits of using unlabelled data for reducing the interval length, especially when labelled data are scarce.

#### **Wasserstein Generative Regression**

*Tong Wang<sup>1</sup>, Shanshan Song<sup>1</sup>, Guohao Shen<sup>2</sup>, Yuanyuan Lin<sup>1</sup> and*

♦*Jian Huang<sup>2</sup>*

<sup>1</sup>The Chinese University of Hong Kong

<sup>2</sup>The Hong Kong Polytechnic University

j.huang@polyu.edu.hk

We propose a Wasserstein generative regression (WGR) approach to learning a general regression function nonparametrically using deep neural networks. WGR is based on an objective function constructed by regularizing the usual least squares loss with the Wasserstein distance between the distribution of the regression function and the data distribution. WGR learns a general regression function that can also serve as a generator for sampling from the conditional distribution of the response given the predictor. This is different from the usual regression methods that only learn the conditional mean or conditional quantile functions of the response given the predictor. Another attractive feature of WGR is that it can easily handle high-dimensional responses and predictors. We present some preliminary results on the consistency and non-asymptotic error bounds of WGR under appropriate conditions. We also conduct extensive numerical experiments to demonstrate the advantages of WGR.

#### **Nonparametric Estimation of Non-Crossing Quantile Regression Process with Deep Requ Neural Networks**

♦*Guohao Shen<sup>1</sup>, Yuling Jiao<sup>2</sup>, Yuanyuan Lin<sup>3</sup>, Joel L. Horowitz<sup>4</sup> and Jian Huang<sup>1</sup>*

## Abstracts

---

<sup>1</sup>The Hong Kong Polytechnic University

<sup>2</sup>Wuhan University

<sup>3</sup>The Chinese University of Hong Kong

<sup>4</sup>Northwestern University

guohao.shen@polyu.edu.hk

We propose a penalized nonparametric approach to estimating the quantile regression process (QRP) in a nonseparable model using rectier quadratic unit (ReLU) activated deep neural networks and introduce a novel penalty function to enforce the non-crossing of quantile regression curves. We establish the non-asymptotic excess risk bounds for the estimated QRP and derive the mean integrated squared error for the estimated QRP under mild smoothness and regularity conditions. To establish these non-asymptotic risk and estimation error bounds, we also develop a new error bound for approximating C $\hat{s}$  smooth functions with  $s > 0$  and their derivatives using ReLU activated neural networks. This is a new approximation result for ReLU networks and is of independent interest and may be useful in other problems. Our numerical experiments demonstrate that the proposed method is competitive with or outperforms two existing methods, including methods using reproducing kernels and random forests for nonparametric quantile regression.

### Sparse Kronecker Product Decomposition: a General Framework of Signal Region Detection in Image Regression

*Sanyou Wu and ♦Long Feng*

University of Hong Kong

lfeng@hku.hk

This paper aims to present the first Frequentist framework on signal region detection in high-resolution and high-order image regression problems. Image data and scalar-on-image regression are intensively studied in recent years. However, most existing studies on such topics focused on outcome prediction, while the research on image region detection is rather limited, even though the latter is often more important. In this paper, we develop a general framework named Sparse Kronecker Product Decomposition (SKPD) to tackle this issue. The SKPD framework is general in the sense that it works for both matrices and (high-order) tensors represented image data. Moreover, unlike many Bayesian approaches, our framework is computationally scalable for high-resolution image problems. Specifically, our framework includes: 1) the one-term SKPD; 2) the multi-term SKPD; and 3) the nonlinear SKPD. We propose nonconvex optimization problems to estimate the one-term and multi-term SKPDs and develop path-following algorithms for the nonconvex optimization. The computed solutions of the path-following algorithm are guaranteed to converge to the truth with a particularly chosen initialization even though the optimization is nonconvex. The nonlinear SKPD is highly connected to shallow convolutional neural networks (CNN), particular to CNN with one convolutional layer and one fully connected layer.

## Session 23CHI110: Recent Advances on Functional Data Analysis

### New Anova Tests for Multivariate Functional Data with Applications

*♦Zhiping Qiu, Jianguan Fan and Jin-Ting Zhang*

zpqiu128@163.com

Multivariate functional data are usually observed in many scientific and technology fields where several continuous vector functions for a statistical unit are obtained on a given interval. This paper is devoted to checking whether the mean vector functions of multi-

sample multivariate functional data are equal. Two new global testing statistics are proposed by integrating or maximizing the pointwise Lawley-Hotelling trace test statistic. The asymptotic expressions of the proposed testing statistics under the null hypothesis are derived, and their root- $n$  consistencies are established. Finally, simulation studies and the applications to two real data examples demonstrate the performance of the two new testing procedures and their advantages over existing testing procedures.

### Long-Memory log-Linear Zero-Inflated Generalized Poisson Autoregression for Covid-19 Pandemic Modeling

*♦Xiaofei Xu<sup>1</sup>, Ying Chen<sup>2</sup>, Yan Liu<sup>3</sup>, Yuichi Goto<sup>4</sup> and Masanobu Taniguchi<sup>3</sup>*

<sup>1</sup>Wuhan University

<sup>2</sup>National University of Singapore

<sup>3</sup>Waseda University

<sup>4</sup>Kyushu University  
xaiofeix@whu.edu.cn

This paper describes the dynamics of daily new cases arising from the Covid-19 pandemic using a long-range dependent model. A new long memory model, LFIGX (Log-linear zero-inflated generalized Poisson integer-valued Fractionally Integrated GARCH process with eXogenous covariates), is proposed to account for count time series data with long-run dependent effect. It provides a novel unified framework for integer-valued processes with serial and long-range dependence (positive or negative), over-dispersion, zero-inflation, nonlinearity, and exogenous variables effects. We adopt an adaptive Bayesian Markov Chain Monte Carlo (MCMC) sampling scheme for parameter estimation. This new modeling is applied to the daily new confirmed cases of Covid-19 pandemic in six countries including Japan, Vietnam, Italy, the United Kingdom, Brazil, and the United States. The LFIGX model provides insightful interpretations on the impacts of policy index and temperature, and delivers good forecasting performance to the dynamics of daily new cases in different countries.

### One-Way Manova for Functional Data via Lawley – hotelling Trace Test

*♦Tianming Zhu<sup>1</sup>, Jin-Ting Zhang<sup>2</sup> and Ming-Yen Cheng<sup>3</sup>*

<sup>1</sup>National Institute of Education, Nanyang Technological University

<sup>2</sup>National University of Singapore

<sup>3</sup>Hong Kong Baptist University  
tianming.zhu@nie.edu.sg

Functional data arise from various fields of study and there have been numerous works on their analysis. However, most of existing methods consider the univariate case and methodology for multivariate functional data analysis is rather limited. In this study, we consider testing equality of vectors of mean functions for multivariate functional data i.e. functional one-way multivariate analysis of variance (MANOVA). To this aim, we study asymptotic null distribution of the functional Lawley–Hotelling trace (FLH) test statistic and approximate it by a Welch–Satterthwaite type chi-squared-approximation. We describe two approaches to estimating the parameters in the chi-squared-approximation ratio-consistently. The resulting FLH test has the correct asymptotic level, is root- $n$  consistent in detecting local alternatives, and is computationally efficient. The numerical performance is examined via some simulation studies and application to two real data examples. The proposed FLH test is comparable with four existing tests based on permutation in terms of size control and power. The major advantage is that it is much faster to compute.

## A Unified Approach to Hypothesis Testing for Functional Linear Models

*Yinan Lin and ♦Zhenhua Lin*

National University of Singapore  
linz@nus.edu.sg

A unified approach to hypothesis testing is developed for scalar-on-function, function-on-function, function-on-scalar models and particularly mixed models that contain both functional and scalar predictors. In contrast with most existing methods that rest on the large-sample distributions of test statistics, the proposed method leverages the technique of bootstrapping max statistics and exploits the variance decay property that is an inherent feature of functional data, to improve the empirical power of tests especially when the sample size is limited or the signal is relatively weak. Theoretical guarantees on the validity and consistency of the proposed test are provided uniformly for a class of test statistics.

## Session 23CHI12: Recent Advances in Variable Selection and Regression Analysis with Interval-Censored Data

### Joint Modelling of Current Status Data and Competing Risks

*♦Da Xu, Tao Hu and Jianguo Sun*

xud062@nenu.edu.cn

Current status data and competing risks data are two types of data that commonly occur in event history studies and many methods have been developed for their analysis separately. Sometimes one may be interested in or need to conduct their joint analysis such as in the clinical trials with composite endpoints. In this paper, we introduce a joint model which combines the competing risks data and current status data. For the problem, the maximum likelihood estimation procedure is derived and in particular, a novel EM algorithm, which is quite stable and can be easily implemented, is developed. Also the asymptotic properties of the resulting estimators are established and the numerical study indicates that the proposed methodology works well in practical situations. An application is provided for the illustration of the presented method.

### Transformation Models with Latent Variables and Informative Partly Interval-Censored Data

*♦Jingjing Jiang<sup>1</sup>, Chunjie Wang<sup>1</sup>, Pan Deng<sup>2</sup> and Xinyuan Song<sup>3</sup>*

<sup>1</sup>Changchun University of Technology

<sup>2</sup>Huazhong University of Science and Technology

<sup>3</sup>The Chinese University of Hong Kong  
jiangjj2123@163.com

This paper considers a new joint modeling approach to simultaneously model the failure and observation times and correlate these two stochastic processes through shared latent factors. The proposed model comprises a transformation model for the failure time of interest, a proportional hazards model for the length of censoring interval, and a factor analysis model for characterization of the latent factors. A multi-stage data augmentation procedure is introduced to tackle the challenges posed by the complex model and data structure. A Bayesian approach coupled with monotone spline approximation and Markov chain Monte Carlo techniques is developed to estimate the unknown parameters and nonparametric functions. The satisfactory performance of the proposed method is demonstrated through simulations, and it is then applied to a Framingham Heart study.

## Variable Selection for Bivariate Interval-Censored Failure Time Data under Linear Transformation Models

*♦Rong Liu<sup>1</sup>, Mingyue Du<sup>2</sup> and Jianguo Sun<sup>3</sup>*

<sup>1</sup>Chnaghcun University of Technology

<sup>2</sup>Jilin University

<sup>3</sup>University of Missouri  
664225997@qq.com

Abstract: Variable selection is needed and performed in almost every field and a large literature on it has been established, especially under the context of linear models or for complete data. Many authors have also investigated the variable selection problem for incomplete data such as right-censored failure time data. In this paper, we discuss variable selection when one faces bivariate interval-censored failure time data arising from a linear transformation model, for which it does not seem to exist an established procedure. For the problem, a penalized maximum likelihood approach is proposed and in particular, a novel Poisson-based EM algorithm is developed for the implementation. The oracle property of the proposed method is established, and the numerical studies suggest that the method works well for practical situations. Keywords: bivariate failure time data; EM algorithm; oracle property; transformation models.

### Simultaneous Variable Selection and Estimation for Interval-Censored Failure Time Data with Ancillary Information

*♦Mingyue Du and Xingqiu Zhao*

The Hong Kong Polytechnic University  
dummoon@163.com

Simultaneous variable selection and estimation has recently attracted a great deal of attention and in particular, many methods have been proposed for it in the context of failure time data. We consider the same problem but for a situation that has been not discussed and when one faces interval-censored data with the presence of some ancillary information given in the form of recurrent event processes. One special case of such situations is interval-censored data with informative censoring. For the problem, a conditional proportional hazards model is proposed and a penalized sieve maximum likelihood procedure is developed for the simultaneous variable selection and estimation of covariate effects. In the method, B-splines functions are used and the oracle property of the proposed estimators is established. A simulation study is conducted to assess the finite sample performance of the proposed approach and suggests that it works well. The method is applied to a set of real data arising from an Alzheimer's disease study.

## Session 23CHI124: Order-of-Addition Experimental Designs

### Design for Order-of-Addition Experiments with Two-Level Components

*Hengzhen Huang*

Guangxi Normal University  
hzhuang@mailbox.gxnu.edu.cn

The statistical design for order-of-addition (OofA) experiments has received much recent interest as its potential in determining the optimal sequence of multiple components, for example, the optimal sequence of drug administration for disease treatment. The traditional OofA experiments focus mainly on the sequence effects of components, i.e., the experimenters fix the factor level of each component and observe how the response is affected by varying the sequences of components. However, the components may also have factorial

## Abstracts

---

effects in that changing their factor levels in a given sequence can affect the response. In view of this, we consider the design problem for OoFA experiments where each component is experimented at two levels. A systematic method is given to construct OoFA designs that jointly considers the sequence design and factorial design for all components. By appropriately choosing the sequence and factorial designs, we show that the combination of the two parts may result in a balanced design with an economical run size. Moreover, the constructed designs enjoy a number of optimality properties such as D-, A- and E-optimalities under some empirical models. The proposed design method can be extended to many other practical situations.

### Fast Approximation of the Shapley Values Based on Order-of-Addition Experimental Designs

♦Liuqing Yang<sup>1</sup>, Yongdao Zhou<sup>1</sup>, Haoda Fu<sup>2</sup>, Min-Qian Liu<sup>1</sup> and Wei Zheng<sup>3</sup>

<sup>1</sup>Nankai University

<sup>2</sup>Eli Lilly and Company

<sup>3</sup>University of Tennessee  
yliuqing0714@163.com

Shapley value is originally a concept in econometrics to fairly distribute both gains and costs to players in a coalition game. In the recent decades, its application has been extended to other areas such as marketing, engineering and machine learning. However, its heavy computational burden has been long recognized but rarely investigated. Specifically, in a  $d$ -player coalition game, calculating a Shapley value requires the evaluation of  $d!$  or  $2^d$  marginal contribution values. Hence it becomes infeasible to calculate the Shapley value when  $d$  is reasonably large. A common remedy is to take a random sample of the permutations to surrogate for the complete list of permutations. We find an advanced sampling scheme can be designed to yield much more accurate estimation of the Shapley value than the simple random sampling (SRS). Our sampling scheme is based on combinatorial structures in the field of design of experiments, particularly the order-of-addition experimental designs for the study of how the orderings of components would affect the output. We show that the obtained estimates are unbiased, and can sometimes deterministically recover the original Shapley value. Both theoretical and simulations results show that our sampling scheme outperforms SRS in terms of estimation accuracy.

### Blocked Designs with Multi Block Variables

♦Shengli Zhao<sup>1</sup>, Qianqian Zhao<sup>1</sup>, Yuna Zhao<sup>2</sup> and Mingqian Liu<sup>3</sup>

<sup>1</sup>Qufu Normal University

<sup>2</sup>Shandong Normal University

<sup>3</sup>Nankai University  
zhaoshengli@qfnu.edu.cn

Fractional factorial designs are commonly used in various experiments. Randomization is one of the three fundamental principles that need to be considered in the design of an experiment, which involves a completely random allocation of the selected treatment combinations to the experimental units. This kind of allocation is appropriate only if the experimental units are homogenous. However, such homogeneity may not always be guaranteed especially when the size of the experiment units is relatively large. A practical design strategy is to partition the experimental units into homogeneous groups, known as blocks, and restrict randomization separately to each block. There are two kinds of blocking problems. One is called the single block variable problem and the other is called the multiple block variables problem. This talk summarizes some of the recent results on constructing optimal blocked designs with multiple block variables under the minimum aberration, clear, and

general minimum lower order confounding.

### Order-of-Addition Experiments on the Adjacency Relationship

Xinran Zhang, Ruonan Zheng, Min-Qian Liu and ♦Jian-Feng Yang  
Nankai University  
jfyang@nankai.edu.cn

With the development of science and technology, the scale of data becomes huge. Design of experiments plays a more and more important role in different areas. In many cases, we focus on the addition order of components, which is the order-of-addition (OoA) problem. The adjacency relationship (AR) between components usually has an important impact on the final results, but there is still very little research on this topic, while most of the existing order-of-addition experiments focus on the relative positions between (or absolute positions of) components. This paper focuses on the issues that the AR between components makes an effect on the response. Starting from the “Traveling Salesman Problem”, this paper proposes an AR model and presents an algorithm for inferring the optimal order. In addition, a loopfull design, which is different from the OoA full design, is proposed. The properties of the loop-full design and the D-optimality under the AR model are given. This paper also puts forward a kind of minimal-point design, which greatly reduces the run size and performs well among the designs with the same run size. Based on the minimal-point design, we bring up an improved algorithm to improve its D-efficiency. The simulation results show

### Session 23CHI146: Recent Developments in Complex Data Structure Analysis

#### An Iterative Model-Free Feature Screening Procedure: Forward Recursive Selection

♦Siwei Xia<sup>1</sup> and Yuehan Yang<sup>2</sup>

<sup>1</sup>School of Science, Civil Aviation Flight University of China

<sup>2</sup>School of Statistics and Mathematics, Central University of Finance and Economics  
xsw@cqu.edu.cn

Many researchers have studied the combinations of machine learning techniques and traditional statistical strategies, and proposed effective procedures for complicated data sets. Yet, there is still some lack of running time and prediction accuracy. In this paper, we propose an iterative feature screening procedure, named forward recursive selection. We combine the random forest and forward selection to address the model-based limitations and the related requirements. We also use the forward strategy with a limited number of iterations to improve the computational efficiency. To provide the theoretical guarantees of this method, we calculate functions of the permutation importance of this algorithm in different models and data with group structures. Numerical comparisons and empirical analysis support our results, and the proposed procedure works well.

#### Modeling Air Quality Level with Autoregression

Fukang Zhu

Jilin University  
zfk8010@163.com

This talk consists two parts. In the first part, inspired by the study of air quality level data, we propose a new model for the normalcy-dominant ordinal time series, which is based on a zero-one-inflated bounded Poisson distribution with an autoregressive feedback mechanism in intensity. Under certain conditions, the stationarity and maximum likelihood estimation are established. Moreover, a Lagrange multiplier test is constructed to detect the

inflation phenomenon. Applications find that the model can adequately capture the air quality level data in 30 major cities in China. In the second part, we propose a novel categorical time series model to study urban air quality, which is based on a linear combination of bounded Poisson distribution and discrete distribution to describe the dynamic and systemic features of air quality, respectively. We estimate the parameters through an adaptive Bayesian Markov chain Monte Carlo sampling scheme and show the satisfactory finite sample performance of the model through simulation studies. Daily air quality level data of Beijing, Shanghai and Guangzhou are analyzed.

### **Dimension Reduction of High-Dimension Categorical Data with Two or Multiple Responses Considering Interactions Between Responses**

*Yuehan Yang*

Central University of Finance and Economics

yyh@cuef.edu.cn

This paper focuses on modeling the categorical data with two or multiple responses. We study the interactions between the responses and propose an efficient iterative procedure based on sufficient dimension reduction. We show that the proposed method reaches the local and global dimension reduction efficiency. The theoretical guarantees of the method are provided under the two- and multiple-response models. We demonstrate the uniqueness of the proposed estimator, further, we prove that the iteration converges to the oracle least squares solution in the first two and  $q$  steps for the two- and multiple-response model, respectively. For data analysis, the proposed method is efficient in the multiple-response model and performs better than some existing methods built in the multiple-response models. We apply this modeling and the proposed method to an adult dataset and a right heart catheterization dataset. Results show that both datasets are suitable for the multiple-response model and the proposed method always performs better than the compared methods.

### **Capped Asymmetric Elastic Net Support Vector Machine for Robust Binary Classification**

♦*Kai Qi and Hu Yang*

College of Mathematics and Statistics, Chongqing University, Chongqing, China  
qkai94@163.com

Recently, there are lots of literature on improving the robustness of SVM by constructing nonconvex functions, but they seldom theoretically study the robust property of the constructed functions. In this paper, based on our recent work, we present a novel capped asymmetric elastic net (CaEN) loss and equip it with the SVM as CaENSVM. We derive the influence function of the estimators of the CaENSVM to theoretically explain the robustness of the proposed method. Our results can be easily extended to other similar nonconvex loss functions. We further show that the influence function of the CaENSVM is bounded, so that the robustness of the CaENSVM can be theoretically explained. Other theoretical analysis demonstrates that the CaENSVM satisfies the Bayes rule and the corresponding generalization error bound based on Rademacher complexity guarantees its good generalization capability. Since CaEN loss is concave, we implement an efficient DC procedure based on the stochastic gradient descent algorithm (Pegasos) to solve the optimization problem. A host of experiments are conducted to verify the effectiveness of our proposed CaENSVM model.

### **Session 23CHI150: Statistical Analysis of Massive Data and Network Data**

#### **Divide-and-Conquer Mcmc for Massive Stationary Time Series Data via Spectral Methods**

*Feng Li*

feng.li@cuef.edu.cn

The temporal dependence inherent in time series models precludes embarrassingly parallel algorithms for inference without imposing independence assumptions that may be unrealistic in practice. We propose an approach that enables the use of efficient cluster-computing frameworks in the frequency domain, where the independence assumption is justified asymptotically in the number of observations. We demonstrate that running an embarrassingly parallel divide-and-conquer algorithm in the time domain results in distorted inference, especially when the time series data has longer memory. In contrast, our proposed method for carrying out embarrassingly parallel computations in the frequency domain gives accurate inference, even in the presence of longer memory. Our approach can be coupled with any embarrassingly parallel divide-and-conquer algorithm. We demonstrate our method for several recently proposed divide-and-conquer Markov chain Monte Carlo algorithms for independent data, thereby extending them to stationary time series data.

#### **A Degree-Corrected Cox Model for Dynamic Networks**

*Yuguo Chen, ♦Lianqiang Qu, Jinfeng Xu, Ting Yan and Yunpeng Zhou*  
qulianq@amss.ac.cn

Continuous time network data have been successfully modeled by multivariate counting processes, in which the intensity function is characterized by covariate information. However, degree heterogeneity has not been incorporated into the model which may lead to large biases for the estimation of homophily effects. In this paper, we propose a degree-corrected Cox network model to simultaneously analyze the dynamic degree heterogeneity and homophily effects for continuous time directed network data. Since each node has individual-specific in- and out-degree effects in the model, the dimension of the unknown time-varying parameters grows with the number of nodes, which makes the estimation problem non-standard. We develop a local estimating equations approach to estimate unknown time-varying parameters, and establish consistency and asymptotic normality of the proposed estimators by using the powerful martingale process theories. We further propose test statistics to test for trend and degree heterogeneity in dynamic networks. Simulation studies are provided to assess the finite sample performance of the proposed method and a real data analysis is used to illustrate its practical utility.

#### **Community Detection of Discrete-Time Temporal Networks under Dynamic Stochastic Block Models**

*Binghui Liu*

liubh100@nenu.edu.cn

Detecting communities of discrete-time temporal networks under dynamic stochastic block models has drawn considerable attention in recent years. The dynamic stochastic block models are established by combining static stochastic block models with Markov chains of community variables, where Markov chains depict evolution of communities variables. Fitting a dynamic stochastic block model by maximizing its likelihood function is a highly nontrivial problem, especially for large-scale networks, as the community variables are both potential and dynamic. To deal with this problem, in this paper, we propose a dynamic profile-pseudo likelihood

## Abstracts

---

method (dPPL) , based on an elaborately designed surrogate of the likelihood function, which has advantages in both computational efficiency and accuracy of community detection, and also has solid theoretical guarantee of convergence and consistency. In addition, we extend it to deal with discrete-time dynamic networks with degree heterogeneity. Finally, we demonstrate the advantages and practicability of the proposed method through extensive simulation results and some real data analyses.

### Grid Point Approximation for Distributed Nonparametric Smoothing and Prediction

♦Yuan Gao<sup>1</sup>, Rui Pan<sup>2</sup>, Feng Li<sup>2</sup>, Riquan Zhang<sup>1</sup> and Hansheng Wang<sup>3</sup>

<sup>1</sup>East China Normal University

<sup>2</sup>Central University of Finance and Economics

<sup>3</sup>Peking University

yuan\_gao96@126.com

Kernel smoothing is a widely used nonparametric method in modern statistical analysis. The problem of efficiently conducting kernel smoothing for a massive dataset on a distributed system is a problem of great importance. In this work, we find that the popularly used one-shot type estimator is highly inefficient for prediction purposes. To this end, we propose a novel grid point approximation (GPA) method, which has the following advantages. First, the resulting GPA estimator is as statistically efficient as the global estimator under mild conditions. Second, it requires no communication and is extremely efficient in terms of computation for prediction. Third, it is applicable to the case where the data are not randomly distributed across different machines. To select a suitable bandwidth, two novel bandwidth selectors are further developed and theoretically supported. Extensive numerical studies are conducted to corroborate our theoretical findings. A real data example is also provided to demonstrate the performance of our GPA method.

## Session 23CHI151: Statistical Learning for Large-Scale and Complex Data

### Robust Distributed Learning

Xiaozhou Wang

East China Normal University

xzwang@sfs.ecnu.edu.cn

The growing size of modern data brings new challenges to many classical statistical problems and calls for the development of distributed learning approaches. While in practice, distributed systems may be attacked or behave abnormally, which causes the distributed algorithms based on faultless systems invalid. We will introduce some research results on robust distributed learning. Algorithms and theoretical properties are given. Simulation studies are conducted to demonstrate the performance of the proposed methodologies.

### Regularized Spectral Clustering under the Mixed Membership Stochasticblock Model

Huan Qing<sup>1</sup> and ♦Jingli Wang<sup>2</sup>

<sup>1</sup>China University of Mining and Technology

<sup>2</sup>Nankai University

jlwang@nankai.edu.cn

Mixed membership community detection is a challenge problem in network analysis. To estimate the memberships under the mixed membership stochastic blockmodels (MMSB), this article proposes two efficient spectral clustering approaches based on regularized Laplacian matrix, Simplex Regularized Spectral Clustering

(SRSC) and Cone Regularized Spectral Clustering (CRSC). SRSC and CRSC methods are developed based on the simplex structure and the cone structure in the variants of the eigen-decomposition of the regularized Laplacian matrix. We show that these two approaches SRSC and CRSC are asymptotically consistent under mild conditions by providing error bounds for the inferred membership vector of each node under MMSB. These two proposed approaches are successfully applied to synthetic and empirical networks with encouraging results compared with some benchmark methods.

## Session 23CHI152: Causal Reinforcement Learning

### Dnet: Distributional Network for Distributional Individualized Treatment Effects

Shikai Luo

ByteDance

sluo198912@163.com

There is a growing interest in developing methods to estimate individualized treatment effects (ITEs) for various real-world applications, such as e-commerce and public health. This paper presents a novel architecture, called DNet, to infer distributional ITEs. DNet can learn the entire outcome distribution for each treatment, whereas most existing methods primarily focus on the conditional average treatment effect and ignore the conditional variance around its expectation. Additionally, our method excels in settings with heavy-tailed outcomes and outperforms state-of-the-art methods in extensive experiments on benchmark and real-world datasets. DNet has also been successfully deployed in a widely used mobile app with millions of daily active users.

### Enhancing Spaced Repetition Scheduling Through Memory Dynamics Modelling and Stochastic Optimization

Junyao Ye

MaiMemo Inc.

jy.ye@maimemo.com

In this talk, I will discuss our published work, which emphasizes advancements in spaced repetition - a mnemonic technique optimizing memorization by scheduling review tasks. We developed an interpretable memory model with Markov property utilizing 220 million memory behavior logs from students, introducing an innovative framework for spaced repetition that cohesively unites memory prediction and schedule optimization. Our system captures memory dynamics and transforms the scheduling optimization into a stochastic shortest path problem, which is then solved via the value iteration method. Experimental results demonstrate substantial improvements: a 64% reduction in error, and a 17% cost reduction in recall rates prediction and schedule optimization. Our proposed system, deployed in the MaiMemo language-learning application, is presently assisting millions of Chinese students. We also constructed and publicly released the first benchmark dataset for spaced repetition, containing invaluable time-series data. The presentation will shed light on the theoretical foundation and methodology of our work.

### Simultaneous Feature Selection and Clustering Based on Square Root Optimization

♦He Jiang<sup>1</sup>, Shihua Luo<sup>2</sup> and Yao Dong<sup>2</sup>

<sup>1</sup>Xi'an Jiao Tong University

<sup>2</sup>Jiangxi University of Finance and Economics

jiangsky2005@126.com

The fused least absolute shrinkage and selection operator (LASSO) simultaneously pursuing the joint sparsity of coefficients and their

successive differences has attracted significant attention for analytics purposes. Although it is extensively used, especially when the number of features exceeds the sample size, tuning the regularization parameters, which depends on noise level, is a challenging task since it is difficult to estimate accurately. To tackle this problem, in this paper, we propose and study square root fused LASSO, which combines the square root loss function and joint penalty functions. In theory, we show that the proposed method can achieve the same error rate as that of fused LASSO by proving its estimation and prediction error bounds. In addition, the error rate of square root fused LASSO is lower than those of LASSO and square root LASSO via simultaneous feature selection and clustering. The choices of the regularization parameters are also shown to be free of . In terms of computation, this work develops a novel algorithm based on the alternating direction method of multipliers algorithm with theoretical guarantee of its convergence. Experiments on simulation and real-world datasets demonstrate the superiority of square root fused LASSO over fused

### Causal Inference and Recommendation System

*Kaixian Yu*

kaixian.yu@shopee.com

The utilization of causal inference in different fields, such as recommendation systems, is a valuable tool. These systems aid users in discovering products, services, or content that align with their interests by providing suggestions that match their preferences. Causal inference offers effective techniques for detecting cause-and-effect associations behind user preferences and is increasingly prevalent in various recommendation scenarios. This presentation will delve into the application of causal inference in e-commerce recommendation systems, using actual business examples to focus on the click-through-rate prediction debiasing problem. We will also address the obstacles and constraints of leveraging causal inference in this setting and offer promising solutions.

### Session 23CHI172: Junior Researcher Award Session

#### Distributional Shift-Aware Off-Policy Interval Estimation

♦Wenzhuo Zhou<sup>1</sup>, Yuhang Li<sup>2</sup>, Ruqiang Zhu<sup>2</sup> and Annie Qu<sup>1</sup>

<sup>1</sup>University of California Irvine

<sup>2</sup>University of Illinois Urbana Champaign  
wenzhuz3@uci.edu

We study high-confidence off-policy evaluation in the context of infinite-horizon Markov decision processes, where the objective is to establish a confidence interval (CI) for the target policy value using only offline data pre-collected from unknown behavior policies. This task faces two primary challenges: providing a comprehensive and rigorous error quantification in CI estimation, and addressing the distributional shift that results from discrepancies between the distribution induced by the target policy and the offline data-generating process. Motivated by an innovative unified error analysis, we jointly quantify the two sources of estimation errors: the misspecification error on modeling marginalized importance weights and the statistical uncertainty due to sampling, within a single interval. This unified framework reveals a previously hidden tradeoff between the errors, which undermines the tightness of the CI. Relying on a carefully designed discriminator function, the proposed estimator achieves a dual purpose: breaking the curse of the tradeoff to attain the tightest possible CI, and adapting the CI to ensure robustness against distributional shifts. The numerical performance of the proposed method is examined in synthetic datasets and an OhioT1DM mobile health study.

#### Probing Differential Expression Patterns Efficiently and Robustly Through Adaptive Linear Multi-Rank Two-Sample Tests

*Dan Daniel Erdmann-Pham*

Stein Fellow, Statistics Department at Stanford University  
erdpham@stanford.edu

Two- and K-sample tests are commonly used tools to extract scientific discoveries from data. Naturally, the precise choice of test ought depend on the specifics of the generating mechanisms producing the data: strong parametric assumptions allow for efficient likelihood-based testing, while non-parametric approaches like Mann-Whitney and Kolmogorov-Smirnov-type tests are popular when such prior knowledge is absent. As this talk will argue, practitioners often find themselves in situations of neither full knowledge of all involved distribution nor in full ignorance of them, and therefore are in need of tests that span the spectrum of possible prior knowledge gracefully. It proposes so-called adaptive linear multi-rank statistics as promising candidates for this task, and illustrates their general utility, flexibility (including applications to multiple testing and testing under nuisance alternatives), and computational feasibility on examples from population genetics and single-cell differential expression analysis.

#### Efficient Algorithms for Large-Scale Optimal Transport Problems

*Cheng Meng*

Renmin University of China  
chengmeng@ruc.edu.cn

Optimal transport methods have played an increasingly predominant role in machine learning, deep learning, statistics, computer vision, and biomedical research. Despite the wide application, existing methods for solving optimal transport problems may suffer from a substantial computational burden when the sample size is large. To alleviate the computational burden, we introduce the Spar-Sink algorithm, which utilizes a novel importance sparsification scheme to accelerate the popular Sinkhorn algorithm. This approach can be effectively applied to the entropic optimal transport problem, unbalanced optimal transport problem, Gromov-Wasserstein distance approximation, Wasserstein barycenter estimation, and generative modeling, among others. We consider the application on echocardiography, which is one of the most promising techniques to display the movements of myocardium. Experiments demonstrate our approach can effectively identify and visualize cardiac cycles, from which one can diagnose heart failure and arrhythmia. To evaluate the numerical accuracy of cardiac cycle prediction, we consider the task of predicting the end-diastole time point using the end-systole one. Results show Spar-Sink performs as well as the classical Sinkhorn algorithm, requiring significantly less computational time.

#### Testing Serial Independence of Object-Valued Time Series

*Feiyu Jiang*

Fudan university  
jiangfy@fudan.edu.cn

We propose a novel method for testing serial independence of object-valued time series in metric spaces, which is more general than Euclidean or Hilbert spaces. The proposed method is fully nonparametric, free of tuning parameters and can capture all nonlinear pairwise dependence. The key concept used in this paper is the distance covariance in metric spaces, which is extended to auto distance covariance for object-valued time series. Furthermore, we propose a generalized spectral density function to account for pairwise dependence at all lags and construct a Cram'er-von Mises type test statistic. New theoretical arguments are developed to establish

## Abstracts

the asymptotic behavior of the test statistic. A wild bootstrap is also introduced to obtain the critical values of the non-pivotal limiting null distribution. Extensive numerical simulations and three real data applications are conducted to illustrate the effectiveness and versatility of our proposed method.

### Session 23CHI35: Non-Euclidean Data Analysis

#### Intrinsic Mave for Spd Matrices and Beyond

Baiyu Chen, Shuang Dai and ♦ Zhou Yu

East China Normal University  
1369959744@qq.com

In this paper, we target the problem of sufficient dimension reduction with symmetric positive definite matrices valued responses. We propose the intrinsic minimum average variance estimation method and the intrinsic outer product gradient method which fully exploit the geometric structure of the Riemannian manifold where responses lie. We present the algorithms for our newly developed methods under the log-Euclidean metric and the log-Cholesky metric. Each of the two metrics is linked to a commutative Lie group structure that transforms our model defined on a manifold into a Euclidean one whose consistency and asymptotic normality are derived. The proposed methods are then naturally extended to general Riemannian manifolds. Several simulation studies and an application to the New York taxi network data are performed to show the superiority of the proposed methods.

#### Testing Strict Stationarity of Complex Time Series

♦Qiang Zhang<sup>1</sup>, Wenliang Pan<sup>2</sup>, Jin Zhu<sup>3</sup> and Xueqin Wang<sup>4</sup>

<sup>1</sup>Chengdu University

<sup>2</sup>Chinese Academy of Sciences

<sup>3</sup>Sun Yat-Sen University

<sup>4</sup>University of Science and Technology of China  
zhangqiang@cdu.edu.cn

We introduce a model-free test for strict stationarity of a sequence of random objects that take values in a general metric space. Our proposed test can identify any deviation from strict stationarity. The test statistic is based solely on metric ranks and is scale-invariant, making it robust to outliers or heavy-tailed data. Results from numerical studies and two real data analyses demonstrate that our method has a reasonable Type I error rate and excellent power, especially when the time series is non-Euclidean or varies smoothly in scales and locations.

#### Fpls-Dc: Functional Partial Least Squares Through Distance Correlation for Imaging Genomics

♦Wenliang Pan<sup>1</sup>, Chuang Li<sup>2</sup>, Tengfei Li<sup>3</sup>, Yue Shan<sup>3</sup>, Yun Li<sup>3</sup> and Hongtu Zhu<sup>3</sup>

<sup>1</sup>Academy of Mathematics and Systems Science, Chinese Academy of Sciences

<sup>2</sup>Sun Yat-sen university

<sup>3</sup>University of North Carolina at Chapel Hill  
panwliang@amss.ac.cn

Motivation: Imaging genetics is the integration of imaging and genetic techniques to investigate the effects of genetic variations on organ (e.g., brain or heart) function or structure in order to better understand their impact on behavior and/or disease phenotypes. Recently, more and more organ-wide imaging endophenotypes have been widely used to detect putative genes for complexly inherited neuropsychiatric and neurodegenerative disorders. However, there are two major challenges associated with the joint analysis of such

organ-wide imaging data with genetic data, including high dimensionality and complex relationships. It motivates us to propose a nonlinear inference framework to partially address some of these challenges. Results: We proposed a functional partial least squares through distance correlation (FPLS-DC) framework with two components to efficiently carry out whole-genome wide analyses of functional phenotypes. The first one is to use the first FPLS-derived base function to reduce the dimensionality of image, while screening genetic markers. The second one is to maximize the distance correlation between genetic markers and projected imaging data, which is a linear combination of the first few FPLS- basis functions, through sequential quadratic programming. We use a gamma approximation to efficiently approximate the null distribution of test statistics. Compared with the existing estimation

#### Ball Impurity: Measuring Heterogeneity in General Metric Spaces

Ting Li

The Hong Kong Polytechnic University  
tingeric.li@polyu.edu.hk

Various domains, such as neuroimaging and network data analysis, have data in complex forms which do not possess a Hilbert structure. We propose ball impurity, a general measure of heterogeneity among complex non-Euclidean objects. Our approach measures the difference between distributions in general metric spaces by generalizing the impurity degree in Hilbert spaces. The measure has properties analogous to triangular inequalities, is straightforward to compute and can be used for variable screening and tree models.

### Session 23CHI38: Recent Advances in Privacy-Protected Data Collection and Analysis

#### Bias Correction in Analysis of Matrix Masked Data

♦Linh Nghiem<sup>1</sup>, Adam Ding<sup>2</sup> and Samuel Wu<sup>3</sup>

<sup>1</sup>University of Sydney

<sup>2</sup>Northwestern University

<sup>3</sup>University of Florida  
linh.nghiem@sydney.edu.au

Recently developed, triple matrix masking is a method that provides a differential privacy guarantee for all parties involved in the data collection process. It is of interest to study the statistical utility of the masked data that was released. In this talk, we examine the effect of masking on some common statistical methods, including linear regression models and contingency tables. Not accounting for masking typically leads to biased estimates of model parameters, so bias correction needs to be applied. We will demonstrate a trade-off between privacy and statistical accuracy of these bias-corrected estimators regarding point estimation, interval estimation, and hypothesis testing.

#### Reducing Noise Level in Differentially Private Data Collection Through Matrix Masking

♦Aidong Adam Ding<sup>1</sup>, Samuel Wu, Guanhong Miao and Shigang Chen

<sup>1</sup>Northeastern University (USA)  
a.ding@northeastern.edu

Differential privacy schemes have been widely adopted in recent years to address issues of data privacy protection. We propose a new Gaussian scheme combining with another data protection technique, called random orthogonal matrix masking, to achieve  $(\epsilon, \delta)$ -differential privacy (DP) in data collection. We prove that the additional matrix masking significantly reduces the rate of noise vari-

ance required in the Gaussian scheme to achieve  $(\epsilon, \delta)$ -DP. Specifically, for fixed number of attributes  $p$ , when  $\epsilon = 0$ ,  $\delta = 0$  and the sample size  $n$  exceeds  $\ln(1/\delta)/\epsilon^2$ , the required additive noise variance to achieve  $(\epsilon, \delta)$ -DP is reduced from  $O(\ln(1/\delta)/\epsilon^2)$  to  $O(1/\delta)$ . With much less noise added, the resulting differential privacy protected pseudo data sets allow much more accurate inferences, thus can significantly improve the scope of application for differential privacy.

### A General Differentially Private Learning Framework for Decentralized Data

Lingchen Kong

lchkong@bjtu.edu.cn

TBC

### Session 23CHI42: Recent Developments in Clinical Trial Design and Data Analysis

#### Meta-Analytic Evaluation of Surrogate Endpoints in Randomized Controlled Trials with Varying Follow-Up Durations and Non-Proportional Hazards

♦Xiaoyu Tang<sup>1</sup> and Ludovic Trinquart<sup>2</sup>

<sup>1</sup>Pfizer, Shanghai, China

<sup>2</sup>Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, USA  
rainie@bu.edu

Valid surrogate endpoints are of great interest for efficient evaluation of novel therapies. With surrogate and true time-to-event endpoints, meta-analytic approaches for surrogacy validation commonly rely on the hazard ratio, ignore that randomized trials possibly contribute to the meta-analysis for different follow-up durations, and that treatment effects and the strength of surrogacy can vary over time. In this context, we introduce a novel two-stage meta-analytic model to evaluate trial-level surrogacy. Our model is based on the difference in restricted mean survival times as a treatment effect measure. We measure trial-level surrogacy by the coefficient of determination at multiple time points. The model borrows strength across data available at multiple time points, does not involve extrapolation, and is valid regardless of the proportional hazard assumption. Simulation studies showed that the estimates of coefficients of determination are unbiased and have high precision in almost all of the scenarios we examined. In two individual patient data meta-analyses in gastric cancer, estimates of coefficients of determination from our model add insights into how the strength of surrogacy changes over time compared to the multi-RCT Clayton survival copula model.

#### Validation of Predictive Analyses for Interim Decisions in Clinical Trials

Lorenzo Trippa

Dana-Farber Cancer Institute

ltrippa@jimmy.harvard.edu

Adaptive clinical trials use algorithms to predict, during the study, patient outcomes and final study results. These predictions trigger interim decisions, such as early discontinuation of the trial, and can change the course of the study. Poor selection of the Prediction Analyses and Interim Decisions (PAID) plan in an adaptive clinical trial can have negative consequences, including the risk of exposing patients to ineffective or toxic treatments. We present an approach that leverages data sets from completed trials to evaluate and compare candidate PAIDs using interpretable validation metrics. The goal is to determine whether and how to incorporate predictions into

major interim decisions in a clinical trial. Candidate PAIDs can differ in several aspects, such as the prediction models used, timing of interim analyses, and potential use of external data sets. To illustrate our approach, we considered a randomized clinical trial in glioblastoma. The study design includes interim futility analyses on the basis of the predictive probability that the final analysis, at the completion of the study, will provide significant evidence of treatment effects. We examined various PAIDs with different levels of complexity to investigate if the use of biomarkers, external data, or novel algorithms improved interim decisions in

#### A Semi-Mechanistic Dose-Finding Design in Oncology using Pharmacokinetic/Pharmacodynamic Modeling

Xiao Su<sup>1</sup>, ♦Yisheng Li<sup>2</sup>, Peter Müller<sup>3</sup>, Chia-Wei Hsu<sup>4</sup>, Haitao Pan<sup>4</sup> and Kim-Anh Do<sup>2</sup>

<sup>1</sup>PlayStation

<sup>2</sup>University of Texas MD Anderson Cancer Center

<sup>3</sup>University of Texas at Austin

<sup>4</sup>St. Jude Children's Research Hospital  
ysli@mdanderson.org

Common phase I dose-finding designs in oncology are either algorithmic or empirical model-based. We propose a new framework for modeling the dose-response relationship, by systematically incorporating pharmacokinetic (PK) data collected in the trial and the hypothesized mechanisms of the drug effects, via dynamic PK and pharmacodynamic (PD) modeling, as well as modeling of the relationship between a latent cumulative pharmacologic PD effect and a binary toxicity outcome. The resulting design is an extension of the existing designs that make use of pre-specified summary PK information. Our simulation studies show, with moderate departure from the hypothesized mechanisms of the drug action, that the performance of the proposed design on average improves upon the common designs, including the continual reassessment method, Bayesian optimal interval design, modified toxicity probability interval method, and a design called PKLOGIT that models the effect of the area under the concentration-time curve (AUC) on toxicity. In case of considerable departure from the underlying drug effect mechanism, the performance of the proposed design is shown to be comparable with the other designs. We illustrate the proposed design by applying it to the setting of a phase I trial of a  $\alpha$ -secretase inhibitor in metastatic or locally advanced solid tumors.

### Session 23CHI45: Recent Advances in Statistical Methodology

#### Connecting Probability Estimation and Statistical Inference for Random Networks

Yichen Qin

University of Cincinnati

qinyn@ucmail.uc.edu

Estimating the probabilities of connections between vertices in a random network using an observed adjacency matrix is an important task for network data analysis. Many existing estimation methods are based on certain assumptions on network structure, which limit their applicability in practice. Without making strong assumptions, we develop an iterative connecting probability estimation method based on neighborhood averaging. Starting at a random initial point or an existing estimate, our method iteratively updates the pairwise vertex distances, the sets of similar vertices, and connecting probabilities to improve the precision of the estimate. We propose a

## Abstracts

---

two-stage neighborhood selection procedure to achieve the trade-off between smoothness of the estimate and the ability to discover local structure. The tuning parameters can be selected by cross-validation. We establish desirable theoretical properties for our method, and further justify its superior performance by comparing with existing methods in simulation and real data analysis.

### Testing of Social Network Dependence Based on Autoregressive Model

♦*Baisuo Jin, Wenyi Li and Xueqin Wang*

jbs@ustc.edu.cn

Social networks have an increasingly significant impact on people's life, and it is particularly important to analyze and study various social networks. Identifying and determining social dependence is the main goal and advantage of social network analysis. We propose an improved spatial autoregression model, in which we introduce the parameter representing social network dependence and a binary indicator representing the susceptibility of different individuals. Based on the proposed model, a score test statistic is constructed to test whether social network dependency exists (whether  $\beta$  is 0), and the asymptotic distribution of the test statistic under the original hypothesis and alternative hypothesis is also given. We performed simulations to evaluate the empirical performance of the proposed test statistics. Finally, the proposed method is applied to the analysis of the dataset of a real-time streaming video platform for video games.

### Online Statistical Inference for Matrix Contextual Bandit

*Qiyu Han, ♦Will Wei Sun and Yichen Zhang*

Purdue

sun244@purdue.edu

Contextual bandit has been widely used for sequential decision-making based on the current contextual information and historical feedback data. In modern applications, such context format can be rich and can often be formulated as a matrix. Moreover, while existing bandit algorithms mainly focused on reward-maximization, less attention has been paid to the statistical inference. To fill in these gaps, in this work we consider a matrix contextual bandit framework where the true model parameter is a low-rank matrix, and propose a fully online procedure to simultaneously make sequential decision-making and conduct statistical inference. The low-rank structure of the model parameter and the adaptivity nature of the data collection process makes this difficult: standard low-rank estimators are not fully online and are biased, while existing inference approaches in bandit algorithms fail to account for the low-rankness and are also biased. To address these, we introduce a new online doubly-debiasing inference procedure to simultaneously handle both sources of bias. In theory, we establish the asymptotic normality of the proposed online doubly-debiased estimator and prove the validity of the constructed confidence interval.

### On Robust Estimation of Hidden Semi-Markov Regime-Switching Model

♦*Shanshan Qin<sup>1</sup>, Zhenni Tan<sup>2</sup> and Yuehua Wu<sup>2</sup>*

<sup>1</sup>Tianjin University of Finance and Economics

<sup>2</sup>York University

qinsslzu@gmail.com

Regime-switching (RS) models illustrate the movements of data sequence switching among different states, an effective way of separating and identifying the dynamics throughout a study period, which receives wide applications in financial markets. In this paper, we propose a novel robust estimation method for hidden semi-Markov regime-switching model (rHSMS). The proposed method

exploits a Huber  $\rho$ -function for M-estimation to relax the normality assumption against the heavy-tailed errors or observations containing outliers. Besides, the rHSMS permits arbitrary sojourn-time distributions of the hidden states process, which overcomes the limits and inflexibility of possible applications of a hidden Markov regime-switching model (HMS). Furthermore, a likelihood-based estimation procedure is developed to estimate the model's parameters. We demonstrate the robustness, flexibility and accuracy of rHSMS by conducting simulation studies on both two-state and three-state RS models under different sojourn-time distributions and several error types or outliers. Finally, we illustrate the validity and outperformance of rHSMS using S&P 500 weekly data and HSI (HangSeng Index) monthly data.

## Session 23CHI5: Innovative Statistical Methods

### Use of Bayesian Hierarchy Model in Pediatric and Rare Disease with Data Extrapolation

*Hao Zhang and ♦Weiya Xu*

Sanofi China

hao16.zhang@sanofi.com

In the development of drugs for rare disease and pediatric indications, the recruitment of participant is often a main challenge and usually we need to establish efficacy evidence with limited number of patients. It is common to utilize the data borrowed from other studies to help estimate the treatment effect. We will introduce the Bayesian hierarchy model to connect the information from clinical trials of different type and of different sources. A comprehensive simulation will be used to illustrate the details and pros as well as cons of the method and a real case will be discussed showing the concerns in practice such as the control of scale and effect sample size.

### The Graphical Approach with a Bonferroni Mixture of Weighted Simes Tests: a Case Study

♦*Wei Zhang and Kaifeng Lu*

Beigene

wei2.zhang@beigene.com

In a non-oncology trial with 2 treatment arms and 1 control arm, we used the Hochberg approach to adjust for multiplicity for the primary endpoint. FDA commented we should include the secondary endpoint in multiplicity adjustment and suggested to use the pooling method or the graphical approach. We then investigated different multiplicity adjustment approaches, such as the pooling method and the graphical approach. The graphical approach with a Bonferroni mixture of weighted Simes was finally chosen to address FDA's comment. This graphical approach is equivalent to the following gatekeeping procedure. First we apply a truncated Hochberg procedure for the primary endpoint with truncation parameter gamma. If both pairwise comparisons on the primary endpoint are significant, then the two pairwise comparisons on the secondary endpoint will be tested using the regular Hochberg procedure at the full alpha. If only one pairwise comparison on the primary endpoint is significant, then the corresponding pairwise comparison on the secondary endpoint will be tested at a reduced significance level of  $(1-\gamma)*\alpha/2$ . This procedure is easy to implement and uniformly more powerful than the corresponding graphical approach based on weighted Bonferroni tests. The protocol was successfully approved by FDA.

### Dealing with Missing Data in Multi-Reader Multi-Case Design

**Studies**

♦*Zhemin Pan<sup>1</sup>, Yingyi Qin<sup>2</sup> and Jia He<sup>2</sup>*

<sup>1</sup>Tongji University

<sup>2</sup>Second Military Medical School

zm\_pan@tongji.edu.cn

Multi-reader multi case (MRMC) studies are commonly used to evaluate the diagnostic accuracies of various imaging modalities, where cases are examined by multiple readers using all available tests. However, missing data can be a common issue in MRMC studies. The traditional approach to handling missing data in MRMC studies is to remove the case, assuming that the missing data is complete at random. This method may lead to a decrease in statistical power. To address this issue, we propose a new method based on the concept of multiple imputation. Our approach combines the mice and MCMC methods with the MRMC analysis method. We evaluated the proposed method through extensive simulations and applied it to a MRMC study of pulmonary nodule detection. Our results demonstrate that the proposed method outperforms the traditional approach in terms of statistical power and accuracy. It provides a more robust and reliable way to handle missing data in MRMC studies. The proposed method has the potential to improve the accuracy and reliability of diagnostic accuracy evaluations in various imaging modalities.

### Quantifying the Investigator's Decision Bias in Clinical Trial: An Approach Based on Causal Mediation Inference

♦*Bo Chen<sup>1</sup>, Xing Zhao<sup>2</sup> and Juying Zhang<sup>2</sup>*

<sup>1</sup>Shenzhen Chipscreen Biosciences CO. LTD; Sichuan University

<sup>2</sup>Sichuan University

chenbof@163.com

The blinded randomized clinical trials are the golden standard. But unblinded investigator is unavoidable sometimes, which will cause the well-known unblinded investigator's bias in the treatment effect. It is natural in such trials, that researchers wish to control the unblinded investigator's bias in analysis stage to correctly recover the treatment effect from a hypothetical trial that investigator was blinded. Under the potential outcome and causal mediation framework, we proposed a novel estimand of quantifying the investigator's decision bias and to recover the blinded trial effect in a study with unblinded investigator. A weighting estimator is also derived, which could provide unbiased investigator's decision bias and recovered blinded treatment effect. Through extensive simulation, we validated that the proposed estimator of investigator's decision bias and the recovered blinded treatment effect are unbiased, respectively. We applied this weighting estimator to a phase III clinical trial to evaluate if there is any unbiased investigator's decision bias exist.

### Session 23CHI78: Recent Developments in Survival Data Analysis

#### Posterior Sampling from the Spiked Models via Diffusion Processes

♦*Yuchen Wu and Andrea Montanari*

Stanford University

wuyc0114@gmail.com

Sampling from the posterior is a key technical problem in Bayesian statistics. Rigorous guarantees are difficult to obtain for Markov Chain Monte Carlo algorithms of common use. In this paper, we study an alternative class of algorithms based on diffusion processes. The diffusion is constructed in such a way that, at its final

time, it approximates the target posterior distribution. The stochastic differential equation that defines this process is discretized (using a Euler scheme) to provide an efficient sampling algorithm. Our construction of the diffusion is based on the notion of observation process and the related idea of stochastic localization. Namely, the diffusion process describes a sample that is conditioned on increasing information. An overlapping family of processes was derived in the machine learning literature via time-reversal. We apply this method to posterior sampling in the high-dimensional symmetric spiked model. We observe a rank-one matrix  $\theta\theta^T$  corrupted by Gaussian noise, and want to sample  $\theta$  from the posterior. Our sampling algorithm makes use of an oracle that computes the posterior expectation of  $\theta$  given the data and the additional observation process. We provide an efficient implementation of this oracle using approximate message passing. We thus develop the first sampling algorithm

#### Model-Free Conditional Screening for Ultrahigh-Dimensional Survival Data via Conditional Distance Correlation

♦*Hengjian Cui<sup>1</sup>, Yanyan Liu<sup>2</sup>, Guangcai Mao<sup>3</sup> and Jing Zhang<sup>4</sup>*

<sup>1</sup>Capital Normal University

<sup>2</sup>Wuhan University

<sup>3</sup>Central China Normal University

<sup>4</sup>Zhongnan University of Economics and Law

jing66@zuel.edu.cn

How to select the active variables that have significant impact on the event of interest is a very important and meaningful problem in the statistical analysis of ultrahigh-dimensional data. In many applications, researchers often know that a certain set of covariates are active variables from some previous investigations and experiences. With the knowledge of the important prior knowledge of active variables, we propose a model-free conditional screening procedure for ultrahigh dimensional survival data based on conditional distance correlation. The proposed procedure can effectively detect the hidden active variables that are jointly important but are weakly correlated with the response. Moreover, it performs well when covariates are strongly correlated with each other. We establish the sure screening property and the ranking consistency of the proposed method and conduct extensive simulation studies, which suggests that the proposed procedure works well for practical situations. Then, we illustrate the new approach through a real dataset from the diffuse large-B-cell lymphoma study.

#### Rank-Based Greedy Model Averaging for High-Dimensional Survival Data

♦*Baihua He<sup>1</sup>, Shuangge Ma<sup>2</sup>, Xinyu Zhang<sup>3</sup> and Lixing Zhu<sup>4</sup>*

<sup>1</sup>International Institute of Finance, School of Management, University of Science and Technology of China

<sup>2</sup>Department of Biostatistics, Yale University

<sup>3</sup>Academy of Mathematics and Systems Science, Chinese Academy of Sciences

<sup>4</sup>Center for Statistics and Data Science, Beijing Normal University at Zhuhai  
baihua@ustc.edu.cn

Model averaging is an effective way to enhance prediction accuracy. However, most previous works focus on low-dimensional settings with completely observed responses. To attain an accurate prediction for the risk effect of survival data with high-dimensional predictors, we propose a novel method: rank-based greedy (RG) model averaging. Specifically, adopting the transformation model with splitting predictors as working models, we doubly use the smooth concordance index function to derive the candidate predictions and op-

## Abstracts

---

timal model weights. The final prediction is achieved by weighted averaging all the candidates. Our approach is flexible, computationally efficient, and robust against model misspecification, as it neither requires the correctness of a joint model nor involves the estimation of the transformation function. We further adopt the greedy algorithm for high dimensions. Theoretically, we derive an asymptotic error bound for the optimal weights under some mild conditions. In addition, the summation of weights assigned to the correct candidate submodels is proven to approach one in probability when there are correct models included among the candidate submodels. Extensive numerical studies are carried out using both simulated and real datasets to show the proposed approach's robust performance compared to the existing regularization approaches.

### Cox Proportional Hazards Cure Model for Incomplete Auxiliary Covariate Data

*Guangcai Mao*

Central China Normal University

maoguangcai@ccnu.edu.cn

TBC

### Session 23CHI82: Advances in Statistical Methods and Inference for Complex Biomedical Data

#### Individualized Dynamic Model for Multi-Resolutinal Data

*Jiuchen Zhang<sup>1</sup>, Fei Xue<sup>2</sup>, Qi Xu<sup>1</sup>, Jung-Ah Lee<sup>1</sup> and ♦Annie Qu<sup>1</sup>*

<sup>1</sup>UCI

<sup>2</sup>Purdue U

aqu2@uci.edu

Mobile health has emerged as a major success in tracking individual health status, due to the popularity and power of smartphones and wearable devices. This has also brought great challenges in handling heterogeneous, multi-resolution data which arise ubiquitously in mobile health due to irregular multivariate measurements collected from individuals. In this paper, we propose an individualized dynamic latent factor model for irregular multi-resolution time series data to interpolate unsampled measurements of time series with low resolution. One major advantage of the proposed method is the capability to integrate multiple irregular time series and multiple subjects by mapping the multi-resolution data to the latent space. In addition, the proposed individualized dynamic latent factor model is applicable to capturing heterogeneous longitudinal information through individualized dynamic latent factors. In theory, we provide the interpolation error bound of the proposed estimator and derive the convergence rate with non-parametric approximation methods. Both the simulation studies and the application to smartwatch data demonstrate the superior performance of the proposed method compared to existing methods.

#### Statistical Inference for Regression Models with a Diverging Number of Covariates: Beyond Linear Regression

*Lu Xia<sup>1</sup>, Bin Nan<sup>2</sup> and ♦Yi Li<sup>3</sup>*

<sup>1</sup>Univ of Washington

<sup>2</sup>UCI

<sup>3</sup>Univ of Michigan  
yili@umich.edu

For statistical inference on regression models with a diverging number of covariates, the existing literature typically makes sparsity assumptions on the inverse of the Fisher information matrix. Such

assumptions, however, are often violated under Cox proportion hazards models, leading to biased estimates with under-coverage confidence intervals. We propose a modified debiased lasso method, which solves a series of quadratic programming problems to approximate the inverse information matrix without posing sparse matrix assumptions. We establish asymptotic results for the estimated regression coefficients when the dimension of covariates diverges with the sample size. As demonstrated by extensive simulations, our proposed method provides consistent estimates and confidence intervals with nominal coverage probabilities. The utility of the method is further demonstrated by assessing the effects of genetic markers on patients' overall survival with the Boston Lung Cancer Survival Cohort, a large-scale epidemiology study investigating mechanisms underlying the lung cancer.

#### Use of Electronic Health Records (Ehr) Data for Research: Challenges and Opportunities

*Hulin Wu*

University of Texas Health Science Center at Houston  
Hulin.Wu@uth.tmc.edu

In this talk, the challenges and opportunities from the real-world EHR data are introduced and discussed from a Big Data perspective. In particular, we propose a 9-step procedure that describes the whole lifecycle of EHR research projects based on our experience: 1) Initiate a project: proposing a research topic with potential high-impact biomedical/clinical questions or hypotheses; 2) Data queries and data extraction; 3) Data cleaning; 4) Data processing; 5) Data preparation; 6) Data analysis, modeling and prediction; 7) Result validation; 8) Result interpretation; and 9) Publication and dissemination. From each of these steps, we will discuss the challenges and opportunities for statisticians. Real data examples for prediction of HIV-infected individuals based a large nationwide EHR database in USA will be used to illustrate the principles and concepts in EHR data processing and analysis. This is a collaboration work with my PhD students, Tianheng Zhang and Yuxuan Gu.

#### Estimating the Reciprocal of a Binomial Proportion

*♦Jiajin Wei<sup>1</sup>, Ping He<sup>2</sup> and Tiejun Tong<sup>1</sup>*

<sup>1</sup>Department of Mathematics, Hong Kong Baptist University

<sup>2</sup>Faculty of Science and Technology, BNU-HKBU United International College  
20482051@life.hkbu.edu.hk

The binomial proportion is a classic parameter with many applications and has been extensively studied in the literature. By contrast, the reciprocal of the binomial proportion, or the inverse binomial proportion, is often overlooked, even though it also plays an important role in various fields. To estimate the inverse binomial proportion, the maximum likelihood method fails to yield a valid estimate when there is no successful event in the Bernoulli trials. To overcome this zero-event problem, several methods have been introduced in the previous literature. Yet to the best of our knowledge, there is little work on a theoretical comparison of the existing point estimators and the interval estimation for the inverse binomial proportion. To fill the gap, we first review some commonly used point estimators, and then develop a new estimator that aims to eliminate the estimation bias. Moreover, we apply four different methods to construct the confidence intervals (CIs) and further study their respective statistical properties. Numerical studies are conducted to evaluate the finite sample performance of the proposed estimators, followed by a recent meta-analysis on the prevalence of heart failure among COVID-19 patients with mortality.



# Index of Authors

- Adams, T, 66, 162  
 Adhikari, S, **54, 123**  
 Aevermann, B, 71, 180  
 Ahn, J, 45, 92  
 Almeida, JRD, 66, 164  
 An, B, **62, 149**  
 Bachoc, F, 71, 178  
 Bae, W, 44, **71, 90, 181**  
 Bai, S, 60, 140  
 Bai, Y, **72, 182**  
 Bai, Z, 60, 141  
 Beaulieu-Jones, B, 50, 109  
 Ben-Michael, E, 63, 153  
 Berg, E, **65, 158**  
 Bian, J, 49, 104  
 Bonzel, C, 50, 109  
 Boodhoo, T, 71, 179  
 Bradley, P, 57, 133  
 Braun, WJ, **67, 167**  
 Buchanan, A, 53, 120  
 Buxbaum, J, 57, 133  
 Buyske, S, 48, 104  
 Cai, G, 51, 113  
 Cai, H, **57, 130**  
 Cai, J, 44, **66, 89, 162**  
 Cai, K, **68, 167**  
 Cai, T, 50, 108, 109  
 Cai, TT, 61, 145  
 Cai, X, **42, 82**  
 Cai, Y, 57, 130  
 Cai, Z, **51, 53, 64, 111, 117, 154**  
 Candes, E, 63, 151  
 Cang, Z, **51, 113**  
 Canhong, W, 53, 118  
 Cao, C, **59, 138**  
 Cao, H, **46, 62, 96, 147**  
 Cao, J, 59, 138  
 Cao, M, **53, 118**  
 Cao, S, 52, 55, 115, 124  
 Cao, X, 40, 77  
 Cao, Y, **45, 94**  
 Chambers, R, 43, 85  
 Chan, G, 54, 121  
 Chan, S, 48, 102  
 Chang, H, **68, 168**  
 Chang, J, 49, **57, 107, 131**  
 Chang, S, **47, 101**  
 Chang, X, 64, 70, 156, 174,  
     175  
 Chatterjee, N, 48, 104  
 Chen, B, **71, 74, 74, 180, 189, 192**  
 Chen, C, **49, 49, 104, 106**  
 Chen, F, 68, **69, 169, 171**  
 Chen, G, **71, 179**  
 Chen, J, 45, 52, 59, **70, 72, 94, 114, 138, 176, 182**  
 Chen, K, 45, 48, **49, 62, 92, 101, 105, 147**  
 Chen, L, **45, 48, 50, 94, 103, 110**  
 Chen, M, 43, 50, 66, 88, 108, 162  
 Chen, P, 65, 159  
 Chen, S, **41, 49, 49, 51, 64, 74, 80, 104, 106, 107, 113, 155, 189**  
 Chen, T, **53, 119**  
 Chen, X, **44, 51, 54, 55, 57, 60, 70, 71, 91, 113, 123, 126, 131, 143, 175, 181**  
 Chen, Y, **41, 47, 49, 66, 72, 73, 79, 100, 101, 104, 161, 183, 186**  
 Chen, Z, **60, 65, 66, 69, 143, 161, 163, 171**  
 Cheng, M, 72, 183  
 Cheng, Q, **50, 110**  
 Cheng, Y, **62, 147**  
 Cheng, Z, 53, 118  
 Choi, J, 44, 90  
 Chowdhury, S, 56, 129  
 Chu, J, **48, 102**  
 Cook, K, **53, 120**  
 Crouse, W, 42, 83  
 Cui, H, 75, 192  
 Cui, X, 67, 166  
 Cui, Y, 46, 51, 62, **63, 72, 96, 111, 148, 154, 182**  
 Dahabreh, I, 54, 120  
 Dai, B, **56, 127**  
 Dai, G, **52, 114**  
 Dai, M, 61, 146  
 Dai, S, 74, 189  
 Dai, W, **46, 60, 63, 96, 140, 151**  
 Dai, Y, 66, 161  
 Daniels, M, **44, 71, 90, 181**  
 Das, S, 49, 106  
 Datta, S, 62, 147  
 Deng, D, **47, 99**  
 Deng, J, 64, 156  
 Deng, K, **48, 103**  
 Deng, M, **48, 103**  
 Deng, P, 72, 184  
 Deng, S, 71, 179  
 Deng, X, 58, 134  
 Deng, Y, 40, 66, 76, 164  
 Dette, H, 43, 86  
 Ding, A, 74, 189  
 Ding, AA, **74, 189**  
 Ding, B, **61, 146**  
 Ding, P, **52, 115**  
 Ding, S, **48, 53, 102, 119**  
 Ding, Y, **55, 58, 123, 136**  
 Do, K, 74, 190  
 Dong, R, 45, 92  
 Dong, Y, 73, 187  
 Du, J, **44, 57, 91, 132**  
 Du, M, **48, 72, 72, 102, 184, 184**  
 Du, P, 67, 167  
 Du, Y, 70, 71, 176, 180  
 Duan, D, 60, 141  
 Duan, F, **51, 113**  
 Duan, K, 48, 102  
 Duan, R, 49, 104  
 Duarte-Salles, T, 49, 104  
 Erdmann-Pham, DD, **73, 188**  
 Esserman, D, 62, 149  
 Fabrizi, E, 43, 85  
 Falconer, T, 49, 104  
 Fan, J, 41, 60, 72, 81, 141, 183  
 Fan, Q, 60, 142  
 Fan, X, **54, 63, 122, 153**  
 Fan, Y, **61, 144**  
 Fang, F, **60, 140**  
 Fang, H, **66, 162**  
 Fang, K, 59, 69, 137, 174  
 Fang, L, 63, 153  
 Fang, X, **64, 154**  
 Feng, H, 60, 141  
 Feng, L, **72, 183**  
 Feng, Y, **43, 85**  
 Feng, Z, 56, 128  
 Forastiere, L, 60, 140  
 Formentini, S, 55, 126  
 Fortune, T, 54, 122  
 Fryzlewicz, P, 70, 177  
 Fu, B, 43, 87  
 Fu, H, **59, 72, 138, 185**  
 Fu, J, 43, 86  
 Fu, JC, 68, 168  
 Fu, L, 45, 92  
 Gai, Y, 55, 126  
 Gang, B, **45, 92**  
 Gang, M, **71, 182**  
 Gao, G, **60, 142**  
 Gao, L, 58, 134  
 Gao, M, 56, 127  
 Gao, T, 51, 111  
 Gao, Y, **58, 71, 73, 134, 178, 187**  
 Gao, Z, **46, 56, 67, 97, 126, 167**  
 Garmire, D, 71, 180  
 Garmire, L, **71, 180**  
 Ge, L, **44, 90**  
 Geng, H, 67, 166  
 Genton, MG, 46, 47, 96, 100  
 Gilbert, P, 40, 76  
 Gong, P, 47, 98  
 Goo, J, 57, 133  
 Goto, Y, 72, 183  
 Gronsbell, J, 50, 109  
 Gu, F, **44, 90**  
 Gu, M, 47, 98  
 Gu, Y, **41, 53, 80, 116**  
 Guan, J, **68, 168**  
 Gui, W, **68, 168**  
 Gunthard, H, 54, 121  
 Guo, B, **47, 98**  
 Guo, F, **60, 141**  
 Guo, J, 46, 97

- Guo, S, **69, 173**  
 Guo, W, **47, 49, 57, 98, 106, 133**  
 Guo, X, **61, 64, 70, 145, 157, 174**  
 Guo, Y, **44, 56, 65, 91, 127, 159**  
 Guo, Z, **60, 141**  
 Hakonarsson, H, **51, 114**  
 Han, D, **52, 67, 115, 166**  
 Han, J, **58, 134**  
 Han, Q, **74, 191**  
 Han, R, **67, 164**  
 Han, X, **62, 150**  
 Han, Z, **58, 58, 135, 135**  
 Hang, W, **64, 154**  
 Hao, M, **66, 164**  
 Hao, W, **57, 131**  
 Hao, Z, **40, 77**  
 Harhay, M, **71, 181**  
 Hayes, R, **45, 92**  
 He, B, **71, 75, 180, 192**  
 He, J, **49, 57, 74, 107, 131, 192**  
 He, K, **46, 95**  
 He, P, **75, 193**  
 He, Q, **57, 133**  
 He, S, **46, 52, 60, 95, 115, 141**  
 He, W, **47, 61, 99, 144**  
 He, X, **42, 45, 55, 63, 83, 91, 125, 152**  
 He, Y, **51, 58, 61, 63, 69, 110, 135, 143, 152, 174**  
 He, Z, **53, 57, 117, 133**  
 Henderson, D, **61, 145**  
 Heping, Z, **53, 118**  
 Hong, S, **61, 145**  
 Horowitz, JL, **72, 182**  
 Hou, C, **60, 141**  
 Hou, J, **50, 109**  
 Hou, L, **43, 46, 66, 86, 94, 163**  
 Hou, Y, **67, 165**  
 Hsu, C, **74, 190**  
 Hsu, L, **54, 57, 121, 133**  
 Hu, C, **66, 161**  
 Hu, F, **65, 160**  
 Hu, G, **57, 131**  
 Hu, J, **45, 56, 60, 92, 126, 141**  
 Hu, L, **54, 54, 70, 123, 123, 177**  
 Hu, T, **72, 184**  
 Hu, X, **41, 43, 81, 87**  
 Hu, Y, **48, 103**  
 Hu, Z, **47, 100**  
 Hua, Z, **59, 138**  
 Huang, B, **57, 132**  
 Huang, C, **53, 59, 119, 137**  
 Huang, D, **41, 51, 64, 68, 69, 79, 111, 156, 169, 173**  
 Huang, H, **41, 49, 72, 82, 105, 184**  
 Huang, J, **46, 48, 52, 64, 66, 67, 72, 72, 96, 101, 157, 163, 164, 182, 182**  
 Huang, L, **42, 85**  
 Huang, Q, **71, 180**  
 Huang, W, **67, 164**  
 Huang, X, **41, 49, 80, 105**  
 Huang, Y, **48, 56, 68, 102, 128, 169**  
 Iberico, L, **53, 120**  
 Imai, K, **63, 153**  
 Ionita-Laza, I, **53, 57, 66, 117, 133, 163**  
 Islam, MN, **49, 104**  
 James, GJ, **45, 92**  
 Ji, J, **64, 155**  
 Ji, P, **57, 130**  
 Jia, B, **50, 108**  
 Jia, R, **59, 70, 138, 175**  
 Jiang, A, **69, 171**  
 Jiang, B, **64, 70, 154, 178**  
 Jiang, D, **66, 68, 161, 170**  
 Jiang, F, **47, 73, 100, 188**  
 Jiang, H, **64, 68, 73, 155, 169, 187**  
 Jiang, J, **61, 72, 145, 184**  
 Jiang, L, **52, 116**  
 Jiang, R, **63, 153**  
 Jiang, W, **43, 54, 86, 122**  
 Jiang, Y, **48, 104**  
 Jiang, Z, **46, 63, 96, 153**  
 Jiao, X, **40, 77**  
 Jiao, Y, **68, 72, 169, 182**  
 Jin, B, **57, 74, 131, 191**  
 Jin, IH, **55, 124**  
 Jin, J, **48, 104**  
 Jin, R, **67, 167**  
 Jin, Y, **63, 68, 151, 170**  
 Jing, B, **46, 63, 97, 153**  
 Johansson, P, **52, 114**  
 Joseph, R, **58, 135**  
 Jui, S, **70, 178**  
 Just, H, **70, 175**  
 Just, HA, **70, 175**  
 Justet, A, **66, 162**  
 Kang, J, **49, 54, 57, 59, 64, 107, 123, 131, 139, 155**  
 Kang, K, **53, 66, 116, 162**  
 Kang, L, **58, 134**  
 Karunamuni, RJ, **61, 146**  
 Kim, C, **49, 104**  
 Kim, J, **58, 134**  
 Kim, JK, **44, 89**  
 Klein, R, **49, 106**  
 Koike, Y, **64, 154**  
 Kollmann, T, **71, 180**  
 Kong, L, **62, 70, 74, 148, 178, 190**  
 Kong, X, **43, 69, 86, 174**  
 Kopetz, S, **52, 116**  
 Kormákksson, M, **60, 142**  
 Kosorok, M, **57, 131**  
 Kotanko, P, **47, 98**  
 Kou, S, **41, 79**  
 Kouyos, R, **54, 121**  
 Lai, J, **62, 150**  
 Lai, M, **62, 147**  
 Lai, T, **62, 148**  
 Lai, TL, **60, 140**  
 Lai, X, **47, 99**  
 Lan, W, **69, 174**  
 Lau, MC, **50, 109**  
 Lawrence, J, **47, 101**  
 Lee, D, **58, 134**  
 Lee, J, **56, 75, 130, 193**  
 Lee, JJ, **55, 123**  
 Lee, M, **45, 93**  
 Lee, MT, **47, 101**  
 Lee, W, **68, 168**  
 Lee, Y, **58, 134**  
 Lei, K, **53, 119**  
 Leng, C, **56, 64, 127, 154**  
 Leng, X, **67, 165**  
 Lesser, G, **69, 171**  
 Levis, A, **54, 120**  
 Li, B, **61, 68, 68, 70, 146, 169, 170, 175**  
 Li, C, **41, 55, 56, 69, 72, 74, 79, 124, 127, 171, 182, 189**  
 Li, D, **47, 54, 55, 71, 100, 121, 126, 179**  
 Li, F, **53, 54, 54, 62, 71, 73, 73, 119, 120, 123, 149, 181, 186, 187**  
 Li, G, **41, 52, 55, 67, 80, 115, 123, 164**  
 Li, H, **45, 47, 92, 100**  
 Li, J, **43, 49, 53, 57, 58, 59, 66, 85, 105, 118, 131, 134, 137, 139, 162**  
 Li, L, **48, 52, 53, 54, 104, 116, 119, 123**  
 Li, M, **54, 57, 60, 62, 70, 70, 123, 133, 141, 149, 175, 176, 177**  
 Li, N, **44, 45, 50, 63, 66, 90, 93, 108, 153, 162**  
 Li, P, **56, 65, 129, 160**  
 Li, R, **67, 71, 165, 180**  
 Li, S, **42, 45, 58, 60, 61, 67, 67, 82, 83, 92, 134, 143, 144, 164, 165**  
 Li, T, **56, 74, 74, 129, 189, 189**  
 Li, W, **51, 56, 65, 69, 70, 74, 110, 127, 160, 171, 172, 177, 191**  
 Li, X, **41, 42, 42, 51, 57, 59, 62, 63, 64, 65, 70, 71, 81, 83, 83, 112, 132, 137, 138, 147, 150, 157, 158, 174, 179**  
 Li, Y, **41, 44, 45, 55-57, 58, 58, 63, 66, 66, 69, 71, 73, 74, 74, 75, 82, 90, 94, 123, 129, 131, 134, 136, 151, 161, 162, 171, 180, 188, 189, 190, 193**  
 Li, Z, **59, 63, 65, 71, 71, 139, 150, 159, 178, 178**  
 Lian, H, **53, 119**  
 Lian, S, **63, 153**  
 Liang, F, **51, 112**  
 Liang, H, **71, 180**  
 Liang, M, **65, 157**  
 Liang, N, **69, 172**  
 Liang, W, **55, 126**  
 Liang, X, **64, 156**  
 Liao, X, **65, 158**  
 Liao, Z, **63, 150**  
 Lim, JCT, **50, 109**  
 Lin, C, **58, 136**  
 Lin, CD, **58, 63, 135, 152**  
 Lin, D, **41, 70, 80, 177**  
 Lin, DKJ, **70, 176**  
 Lin, G, **45, 93**  
 Lin, H, **40, 78**  
 Lin, J, **66, 161**  
 Lin, Q, **68, 169**  
 Lin, R, **55, 124**  
 Lin, X, **51, 52, 63, 114, 116, 150**  
 LIN, Y, **60, 143**  
 Lin, Y, **48, 50, 50, 60, 67, 72, 101, 108, 109, 142, 164, 182, 184**  
 Lin, Z, **42, 43, 49, 59, 71,**

- 72, 84, 87, 107,  
138, **178, 184**
- Linero, A, **71, 181**
- Liu, B, **54, 56, 73, 121, 128,**  
**186**
- Liu, C, **51, 112**
- Liu, F, **43, 88**
- Liu, H, 54, **69, 70, 122, 174,**  
176
- Liu, J, **40, 49, 50, 52, 61,**  
66, **76, 106, 110,**  
146, 163
- Liu, L, 44, **47, 50, 51, 53,**  
57, **64, 90, 99,**  
109, 110, 117,  
133, **154**
- Liu, M, 55, 57, 62, 63, 70,  
72, 73, 125, 133,  
147, 151, 177,  
185
- Liu, P, **57, 69, 130, 172**
- Liu, R, **51, 53, 72, 112, 119,**  
**184**
- Liu, S, **57, 70, 133, 175**
- Liu, T, 52, 114
- Liu, W, 42, 44, 50, **59, 82,**  
89, 109, **137**
- Liu, X, **43, 43, 46, 50, 51,**  
56, **86, 86, 96,**  
110, 113, 127
- Liu, Y, **41, 44, 50, 55,**  
56, 56, 57, **62,**  
63, 63, 64, **65,**  
66, 70, 72, 75,  
**80, 89, 110, 125,**  
128, **129, 133,**  
**150, 151, 153,**  
153, 157, **160,**  
162, 178, 183,  
192
- Liu, Z, **60, 63, 65, 70, 141,**  
151, 152, **160,**  
**175**
- Liu, Y, 46, 96
- Lok, J, **53, 120**
- Long, Q, 49, 106
- Lou, Y, 40, 77
- Lou, Z, 41, 81
- Lu, H, 46, 96
- Lu, J, **44, 50, 60, 62, 88, 89,**  
**109, 141, 147**
- Lu, K, 50, 74, 108, 191
- Lu, T, **57, 132**
- Lu, W, 53, 54, 57, 120, 121,  
131
- Lu, X, **56, 61, 68, 68, 70,**  
**127, 130, 144,**  
**167, 168, 176**
- Lu, Z, 56, 129
- Luan, Y, 52, 114
- Luice, T, 59, 138
- Luo, C, **49, 104**
- Luo, K, 42, 83
- Luo, L, 67, 164
- Luo, S, **73, 73, 187, 187**
- Luo, W, **68, 170**
- Luo, X, **46, 48, 94, 103**
- Luo, Y, 62, 148
- Lyu, Q, 47, 99
- Ma, C, **61, 145**
- Ma, G, 58, 134
- Ma, H, **58, 136**
- Ma, N, 54, 122
- Ma, R, 54, 122
- Ma, S, **46, 50, 53, 58, 59,**  
66, 70, 75, **95,**  
110, **117, 136–**  
138, 162, 174,  
192
- Ma, T, 49, 104
- Ma, W, 43, **65, 87, 160**
- Ma, X, 47, 98
- Ma, Y, 50, 65, 110, 158
- Ma, Z, **66, 162**
- Madrid, O, 56, 127
- Mahmoudi, F, 68, 168
- Mandava, A, 71, 180
- Mao, G, **75, 75, 192, 193**
- Mao, J, 46, 97
- Mao, X, 42, **44, 44, 82, 89,**  
89
- Marder, K, 53, 117
- Mei, H, **59, 138**
- Meng, C, **73, 188**
- Meng, J, 49, 106
- Meng, Q, 59, 138
- Mi, G, 58, 133
- Miao, G, 74, 189
- Miao, W, **55, 59, 63, 66,**  
**125, 137, 154,**  
161
- Miao, Y, **61, 146**
- Ming, H, 53, 118
- Mo, W, **56, 128**
- Montanari, A, 75, 192
- Mu, R, **42, 85**
- Muehlmann, C, 71, 178
- Mukherjee, A, 69, 172
- Müller, P, 74, 190
- Nan, B, 75, 193
- Negash, A, 58, 135
- Ng, CCY, 50, 109
- Ng, MH, 46, 94
- Nghiem, L, **74, 189**
- Nguyen, T, 47, 99
- Ni, L, **51, 112**
- Ni, Q, 61, 145
- Nian, G, **50, 108**
- Nie, L, 68, 170
- Ning, J, **40, 78**
- Ning, W, **42, 42, 83, 83**
- Ning, Y, 65, 157
- Niu, H, 54, 122
- Nordhausen, K, 71, 178
- O'sullivan, F, 44, 90
- Oganisian, A, **71, 181**
- Ohlssen, D, 60, 142
- Others, A, 50, 109
- Ou, Z, **40, 78**
- Ouyang, M, 66, 162
- O'connell, J, 48, 104
- Pai, J, 69, 171
- Pan, B, 70, 178
- Pan, G, 62, 150
- Pan, H, **59, 74, 137, 190**
- Pan, J, 54, 121
- Pan, Q, **65, 158**
- Pan, R, 73, 187
- Pan, T, 57, 131
- Pan, W, **42, 55, 62, 74,**  
74, **84, 124, 149,**  
**189, 189**
- Pan, Y, **45, 93**
- Pan, Z, **74, 192**
- Pang, S, **55, 126**
- Panjwani, N, 52, 116
- Park, C, 54, 120
- Pei, H, 69, 171
- Peligrad, M, 54, 122
- Peng, H, 43, 61, 87, 146
- Peng, J, **56, 129**
- Peng, L, 61, **62, 67, 67, 144,**  
**148, 164, 165,**  
165
- Peng, M, **40, 47, 77, 101**
- Peng, P, **49, 107**
- Perri, M, 71, 181
- Ping, Y, 60, 143
- Pu, D, 69, 174
- Pu, H, 63, 154
- Qi, D, **69, 172**
- Qi, H, **64, 156**
- Qi, J, 66, 162
- Qi, K, **73, 186**
- Qi, L, **63, 152**
- Qi, Y, **62, 149**
- Qi, Z, 51, 70, 111, 177
- Qian, S, **42, 83**
- Qian, Y, 61, 71, 145, 180
- Qiao, G, 59, 138
- Qiao, H, 71, 179
- Qiao, N, 58, 136
- Qiao, X, 56, 128
- Qiao, Y, 42, 82
- Qiao, Z, 57, 130
- Qin, F, 51, 113
- Qin, G, **47, 99**
- Qin, H, 40, 78
- Qin, J, 56, 58, 59, 68, 69,  
129, 136, 137,  
168, 172
- Qin, L, 62, 149
- Qin, S, **74, 191**
- Qin, X, **49, 106**
- Qin, Y, **74, 74, 190, 192**
- Qin, Z, **45, 48, 92, 103**
- Qing, H, 73, 187
- Qiu, B, 45, 94
- Qiu, P, **44, 69, 91, 172**
- Qiu, R, 68, 169
- Qiu, X, **71, 180**
- Qiu, Y, **70, 71, 175, 181**
- Qiu, Z, **72, 183**
- Qu, A, 43, 55, 58, 73,  
75, 88, 126, 136,  
188, **193**
- Qu, L, **73, 186**
- Qu, Z, 46, 61, 96, 144
- Ranalli, M, 43, 85
- Randolph, T, 48, 104
- Ray, D, 58, 135
- Ren, H, **67, 166**
- Ren, M, 53, 118
- Ren, Y, 50, **71, 110, 178**
- Reps, J, 49, 104
- Rinaldo, A, 56, 127
- Robberson, J, 67, 167
- Rubin, D, 41, 79
- Salvati, N, 43, 85
- Sang, H, **54, 54, 122, 122**
- Sang, P, **62, 148**
- Satten, G, 48, 103
- Satter, F, 43, 85
- Scheuermann, R, 71, 180
- Sechidis, K, 60, 142
- Segal, L, 45, 92
- Shan, G, **42, 84**
- Shan, N, **43, 86**
- Shan, Y, 74, 189
- Shang, L, **51, 113**
- Shao, L, 66, 161
- Shao, M, 41, 80
- Shao, Q, 54, 122
- Sharpton, T, 66, 161
- She, R, 60, 141
- She, Y, **51, 53, 112, 119**
- Shen, C, 48, **49, 103, 106**
- Shen, D, **48, 102**
- Shen, G, 48, **72, 72, 101,**  
**182, 182**
- Shen, H, **68, 68, 167, 167**
- Shen, L, 62, 149
- Shen, S, 46, 95
- Shen, T, **51, 111**
- Shen, W, 57, 131
- Shen, X, 51, 55, 113, 124
- Shen, Y, **41, 82**

- Sheng, Y, **59**, **137**  
 Shi, C, 51, 56, 70, 111, 129,  
   177  
 Shi, G, 50, 108  
 Shi, H, 61, 145  
 Shi, J, **46**, **98**  
 Shi, L, 65, 159  
 Shi, S, 47, 99  
 Shi, X, 46, 49, 50, 63, 96,  
   104, 109, 110,  
   154  
 Shi, Y, **45**, **94**  
 Shpitser, I, 66, 161  
 Shu, H, **61**, **144**  
 Shuai, K, 51, 110  
 Si, Y, **49**, **107**  
 Sinha, D, 53, 119  
 Small, D, 54, 120  
 Smith, M, 65, 157  
 Song, C, **52**, **116**  
 Song, F, 52, **62**, 114, **149**  
 Song, H, 60, 141  
 Song, P, 57, 131  
 Song, S, 43, 72, 86, 182  
 Song, X, 40, 43, **48**, 60, 66,  
   67, 72, 77, 85,  
   **101**, 142, 162,  
   165, 184  
 Song, Y, 54, 63, 121, 151  
 Song, Z, 42, **69**, 85, **172**  
 Spiegelman, D, 53, 120  
 Stephens, M, 42, 83  
 Stevenson, I, 60, 142  
 Strug, L, 52, 116  
 Su, B, **43**, **87**  
 Su, W, **54**, **121**  
 Su, X, 74, 190  
 Sun, C, 71, 179  
 Sun, D, **53**, 62, 71, **117**, 147,  
   180  
 Sun, F, **40**, 58, 70, **78**, 135,  
   177  
 Sun, H, 71, 180  
 Sun, I, **41**, **82**  
 Sun, J, 40, **43**, 44, 60, 71,  
   72, 77, **87**, 90,  
   143, 181, 184  
 Sun, K, 70, 178  
 Sun, L, 42, 50, **51**, 52, 84,  
   109, **112**, 116  
 Sun, P, 49, 106  
 Sun, Q, 43, 87  
 Sun, S, 41, 79  
 Sun, T, **58**, **59**, **136**, **139**  
 Sun, W, 45, 57, **59**, 92, 133,  
   **136**  
 Sun, WW, **74**, **191**  
 Sun, Y, **40**, 46, **76**, 95  
 Sun, Z, 62, 147  
 Tada, K, 57, 133  
 Tan, F, **64**, **157**  
 Tan, S, 67, 164  
 Tan, W, 50, 108  
 Tan, Y, 70, 175  
 Tan, Z, 74, 191  
 Tang, B, 44, 91  
 Tang, H, **60**, **140**  
 Tang, L, **47**, **99**  
 Tang, M, 69, 172  
 Tang, N, 43, 88  
 Tang, R, **59**, **139**  
 Tang, S, 59, 138  
 Tang, W, 60, 141  
 Tang, X, **74**, **190**  
 Tang, Y, **40**, **77**  
 Taniguchi, M, 72, 183  
 Tao, L, **51**, 53, **112**, 119  
 Tchetgen, ET, 63, 66, 154,  
   161  
 Teng, Z, **59**, **138**  
 Thall, PF, 52, 116  
 Tian, G, 69, 172  
 Tian, M, **68**, **168**  
 Tian, S, 68, 169  
 Tian, T, **51**, **114**  
 Tian, W, **42**, **83**  
 Tian, X, 62, 149  
 Tian, Y, **63**, 70, **151**, **152**,  
   177  
 Tong, G, **71**, **181**  
 Tong, P, **64**, **155**  
 Tong, T, 46, 47, 75, 97, 98,  
   100, 193  
 Tong, X, **41**, 48, **79**, 102  
 Trapani, L, 69, 174  
 Trinquart, L, 74, 190  
 Trippa, L, **74**, **190**  
 Tsang, KW, **115**  
 Tsui, K, 59, 138  
 Tsui, KL, 59, 137  
 Tsung, F, 47, 67, 99, 166,  
   167  
 Tu, D, 61, 143  
 Tu, J, 44, 89  
 Tudorascu, D, 61, 144  
 Um, S, 54, 123  
 Urquhart, A, 48, 102  
 Usvyat, L, 47, 98  
 Vallejo, J, 68, 170  
 Wan, C, **67**, **166**  
 Wan, H, 48, 103  
 Wan, X, 43, 87  
 Wang, A, **68**, **170**  
 Wang, B, 46, 47, 54, 62, 97,  
   98, 120, 147  
 Wang, C, 40, 44, **45**, 52,  
   **56**, 57, 61, 66,  
   **70**, 72, 77, 88,  
   89, **92**, 116, **129**,  
   133, 146, 163,  
   **177**, 184  
 Wang, D, **47**, 47, 60, **63**,  
   63, **98**, 100, 143,  
   **151**, 151  
 Wang, F, 41, **51**, **52**, 56, 64,  
   70, 71, 79, **111**,  
   **116**, 127, 156,  
   176, 178  
 Wang, G, **54**, 62, **65**, **120**,  
   147, **159**  
 Wang, H, **41**, **42**, 46, 48, 51,  
   56, 58, 59, 64,  
   66, 69–71, 73,  
   **80**, **82**, 95, 104,  
   111, 112, 127,  
   134, 137, 138,  
   156, 161, 173,  
   175, 176, 178,  
   179, 187  
 Wang, J, 45, 47, **51**, **52**, **55**,  
   55, **61**, 64, **66**,  
   69, **70**, **73**, 91,  
   100, **111**, 114,  
   125, **126**, **146**,  
   157, **161**, 172,  
   **177**, 187  
 Wang, L, 40, **41**, **48**, **54**, 62,  
   78, **81**, **101**, **121**,  
   147  
 Wang, M, 43, **49**, **53**, 58, 87,  
   **106**, **118**, 135  
 Wang, N, 66, 162  
 Wang, P, **40**, **46**, **56**, 56, **77**,  
   **97**, **129**, 129  
 Wang, Q, **45**, 58, **59**, 63, **92**,  
   133, **137**, 153  
 Wang, R, 53, 54, 59, 60,  
   119–121, 137,  
   141  
 Wang, S, **40**, 40, **43**, 48, **51**,  
   **70**, **77**, 78, **85**,  
   103, **113**, **177**  
 Wang, T, 56, 65, **66**, **68**, 72,  
   126, 159, **163**,  
   **168**, 182  
 Wang, W, **41**, **44**, **46**, **57**, **81**,  
   **89**, **95**, **132**  
 Wang, X, 44, **46**, **52**, 55, **56**,  
   58, **60**, 61, **62**,  
   62, 66, **67**, 70,  
   71, **73**, 74, 88,  
   89, 91, **97**, **114**,  
   **126**, 126, 134,  
   142, 144, 149,  
   **150**, 164, **166**,  
   176, 182, **187**,  
   189, 191  
 Wang, Y, **40**, 43, 44, **47**–**49**,  
   53, 57, **59**, 61,  
   62, 64, 70, **76**,  
   87, 89, **98**, **104**,  
   **105**, 117, 130,  
   **139**, 144, 147,  
   148, 156, 178  
 Wang, Z, 43, **44**, 50, 66, 71,  
   86, **89**, 108, 162,  
   182  
 Want, J, **44**, **89**  
 Webber, G, 50, 109  
 Wei, G, 60, 142  
 Wei, J, **75**, **193**  
 Wei, Q, **55**, **125**  
 Wei, Y, **43**, 46, 62, 66, **88**,  
   94, 149, 163  
 Wei, Z, 51, 114  
 Weiss, D, 44, 88  
 Wen, W, 60, 143  
 Wen, X, 67, 167  
 Windmeijer, F, 60, 142  
 Wojcik, G, 48, 104  
 Wong, WK, 43, 86  
 Wu, B, **64**, **155**  
 Wu, C, 65, **69**, 159, **172**  
 Wu, F, **64**, **155**  
 Wu, H, **60**, **75**, **141**, **193**  
 Wu, J, 61, **66**, 66, 146, **161**,  
   162  
 Wu, K, 64, 155  
 Wu, M, 49, 57, 107, 131,  
   133  
 Wu, P, **52**, **115**  
 Wu, Q, 41, 44, 48, **69**, 80,  
   90, 102, **171**  
 Wu, R, **61**, **145**  
 Wu, S, 46, 51, **69**, 72,  
   74, 95, 111, **173**,  
   183, 189  
 Wu, T, 57, 130  
 Wu, W, **58**, **135**  
 Wu, X, 53, 119  
 Wu, Y, 64, 70, 74, **75**, 157,  
   175, 191, **192**  
 Wu, Z, 70, 177  
 Xi, R, **66**, **163**  
 Xia, D, 41, 80  
 Xia, L, 41, **50**, 75, 82, **108**,  
   193  
 Xia, S, **73**, **185**  
 Xia, X, **46**, 51, **53**, **94**, 111,  
   118  
 Xia, Y, 64, 154  
 Xia, Z, **61**, **143**  
 Xian, J, 67, 166  
 Xiang, D, **47**, **99**  
 Xiang, X, **65**, **158**  
 Xiao, F, **51**, **113**  
 Xiao, L, **53**, **119**  
 Xiao, Q, **58**, **135**  
 Xiao, R, 62, **71**, 148, **179**

|  |  |  |   |
|--|--|--|---|
| Xiao, X, 61, 146   | Yang, S, <b>58, 67, 134, 165</b>   | Zeng, D, 41, 53, 64, 80, 117,<br>156   | <b>78, 80, 88, 93,</b><br>102, <b>124, 136,</b><br><b>165, 168, 173,</b><br>176, 180, 191   |
| Xiao, Y, 40, 54, <b>57, 71, 78,</b><br>122, <b>132, 180</b>  | Yang, T, 55, 125   | Zeng, H, 51, 111   | Zhang, YD, <b>48, 104</b>   |
| Xie, F, 46, 63, 97, 150                                      | Yang, W, 61, <b>64, 145, 157</b>   | Zeng, P, <b>46, 96</b>   | Zhang, Z, 62, 63, <b>69, 150,</b><br>153, <b>172</b>  |
| Xie, M, <b>44, 90</b>  | Yang, X, 43, 59, 65, <b>71, 87,</b><br>139, 159, <b>180</b>  | Zeng, Q, 50, <b>64, 109, 155</b>   | Zhao, B, 43, 59, 87, 139  |
| Xie, S, <b>64, 156</b>                                       | Yang, Y, <b>41, 45, 50, 54, 57,</b><br><b>62, 67, 70, 73,</b><br>73, <b>82, 94, 108,</b><br>121, <b>133, 147,</b><br><b>166, 175, 185,</b><br><b>186</b> | Zhan, H, 64, 155   | Zhao, H, 43, 44, <b>61, 87, 90,</b><br><b>145</b>   |
| Xie, SX, 45, 94  | Yao, B, 41, 81   | Zhan, J, 48, 104   | Zhao, J, <b>41, 43, 56, 63, 66,</b><br><b>78, 87, 127, 153,</b><br>162  |
| Xie, X, 44, 91   | Yao, D, <b>63, 150</b>   | Zhan, X, 40, 42, <b>45, 76, 85,</b><br><b>93</b>   | Zhao, K, 62, 150  |
| Xie, Y, 43, 86   | Yao, F, <b>55, 67, 124, 166</b>  | Zhang, A, 51, 112  | Zhao, L, <b>58, 59, 133, 139</b>  |
| Xie, Z, 71, 179  | Yao, Q, 46, 64, 97, 154  | Zhang, B, 46, 52, 56, 62, 64,<br>69, 95, 114, 126,<br>149, 156, 173  | Zhao, P, <b>65, 67, 159, 166</b>  |
| Xin, K, <b>41, 82</b>  | Yao, Y, 61, 146  | Zhang, C, <b>41, 49, 55, 59, 60,</b><br>69, <b>79, 105, 124,</b><br><b>139, 142, 172</b>   | Zhao, Q, 72, 185  |
| Xing, C, <b>60, 140</b>                                      | Ye, C, 48, 104   | Zhang, D, 62, 149  | Zhao, R, 48, 50, 54, 55, 104,<br>109, 122, 125,<br>126  |
| Xing, L, 40, 77  | Ye, J, <b>73, 187</b>  | Zhang, H, <b>42, 46, 48, 55, 56,</b><br>66, <b>74, 84, 97,</b><br>104, 126, 163,<br>191  | Zhao, S, 42, 53, 59, <b>72, 83,</b><br>118, 139, <b>185</b>   |
| Xiong, J, 47, <b>61, 99, 144</b>                             | Ye, K, <b>58, 135</b>  | Zhang, J, <b>44, 45, 47, 47–49,</b><br>51, <b>52, 62, 65,</b><br>72, <b>74, 75, 75,</b><br>88, <b>89, 91, 99,</b><br><b>100, 104, 105,</b><br><b>114, 114, 149,</b><br><b>159, 183, 192,</b><br>192, 193 | Zhao, W, 50, 108  |
| Xiong, S, <b>63, 151</b>                                     | Ye, Z, 49, 104   | Zhang, L, 42, 45, <b>53, 61,</b><br>62, <b>63, 65, 65,</b><br>85, 92, <b>118, 145,</b><br>150, <b>153, 160,</b><br>160   | Zhao, X, 44, <b>48, 62, 72,</b><br>74, 89, <b>102, 147,</b><br>184, 192   |
| Xiong, Y, <b>54, 121</b>                                     | Yeong, J, 50, 109  | Zhang, M, 52, 115  | Zhao, Y, <b>42, 42, 43, 47, 48,</b><br>59, <b>62, 65, 66,</b><br>68, <b>70, 70, 72,</b><br><b>84, 85, 85, 100,</b><br>102, 137, <b>149,</b><br>157, <b>163, 169,</b><br>170, <b>176, 176,</b><br>178, 185 |
| Xu, D, 40, <b>72, 77, 184</b>                                | Yi, G, <b>40, 40, 76, 77</b>   | Zhang, N, <b>65, 158</b>   | Zhao, Z, 49, 107  |
| Xu, F, <b>54, 122</b>  | Yi, GY, 57, 132  | Zhang, P, 54, 121  | Zheng, Q, 56, 129   |
| Xu, G, 50, 66, 110, 162                                      | Yi, M, <b>71, 178</b>  | Zhang, Q, 42, 46, <b>57, 59, 68,</b><br>69, <b>74, 83, 95,</b><br><b>132, 137, 169,</b><br>174, <b>189</b>   | Zheng, R, 73, 185   |
| Xu, H, 49, 63, 104, 152                                      | Yi, S, <b>63, 151</b>  | Zhang, R, <b>49, 57, 69, 73,</b><br><b>105, 133, 174,</b><br>187   | Zheng, S, 62, 67, 150, 165  |
| Xu, J, 41, 52, 73, 80, 114,<br>115, 186                      | Yi, Y, <b>42, 50, 85, 109</b>  | Zhang, S, 46, 50, <b>53, 56,</b><br>67, 96, 109, <b>117,</b><br><b>130, 166</b>  | Zheng, W, 43, <b>58, 72, 86,</b><br>134, 185  |
| Xu, K, <b>70, 176</b>  | Yin, X, 50, 58, 109, 135   | Zhang, T, <b>60, 61, 140, 144</b>  | Zheng, X, <b>53, 55, 118, 123</b>   |
| Xu, L, 46, 97  | Ying, A, 62, 148   | Zhang, W, 57, 58, 66, <b>74,</b><br>130, 136, 161,<br><b>191</b>   | Zheng, Y, 67, 164   |
| Xu, M, 44, 89  | Yip, KY, 62, 149   | Zhang, X, <b>40–42, 42, 44,</b><br>46, <b>50, 50, 55–</b><br>58, 73, 75, <b>77,</b><br><b>81, 82, 82, 88,</b><br>96, 98, <b>110, 110,</b><br>126, 128, 132,<br>134, 185, 192                             | Zheng, Z, <b>45, 92</b>   |
| Xu, Q, 43, 58, 75, 88, 136,<br>193                           | You, S, 40, 77   | Zhang, Y, <b>40, 41, 43, 45, 48,</b><br>55, <b>58, 67, 68,</b><br><b>69, 70, 71, 74,</b>   | Zhong, J, <b>56, 128</b>  |
| Xu, R, <b>62, 148</b>  | Yu, B, <b>70, 176</b>  | Zhang, M, 52, 115  | Zhong, W, <b>51, 67, 71, 111,</b><br>165, 182   |
| Xu, S, <b>50, 58, 108, 135</b>                               | Yu, C, <b>47, 58, 100, 134</b>   | Zhang, N, <b>65, 158</b>   | Zhong, Y, 59, 139   |
| Xu, W, 45, <b>47, 66, 74, 92,</b><br><b>100, 164, 191</b>    | Yu, G, <b>56, 128</b>  | Zhang, P, 54, 121  | Zhong, Z, 52, 114   |
| Xu, X, 51, <b>72, 111, 183</b>                               | Yu, H, 41, 82  | Zhang, Q, 42, 46, <b>57, 59, 68,</b><br>69, <b>74, 83, 95,</b><br><b>132, 137, 169,</b><br>174, <b>189</b>   | Zhou, B, 45, 92   |
| Xu, Y, 46, <b>59, 63, 95, 137,</b><br>150                    | Yu, K, <b>73, 188</b>  | Zhang, R, <b>49, 57, 69, 73,</b><br><b>105, 133, 174,</b><br>187   | Zhou, D, 50, 109  |
| Xue, F, 58, 75, 136, 193                                     | Yu, L, 52, 66, 69, 163, 174  | Zhang, S, 46, 50, <b>53, 56,</b><br>67, 96, 109, <b>117,</b><br><b>130, 166</b>  | Zhou, F, <b>70, 70, 177, 178</b>  |
| Xue, H, 42, 84   | Yu, M, 41, <b>67, 81, 167</b>  | Zhang, T, <b>60, 61, 140, 144</b>  | Zhou, H, 44, 89   |
| Xue, L, 68, <b>69, 169, 171</b>                              | Yu, S, 50, 62, 108, 147  | Zhang, W, 57, 58, 66, <b>74,</b><br>130, 136, 161,<br><b>191</b>   | Zhou, J, 44, 90   |
| Xueqin, W, 53, 118   | Yu, T, <b>60, 141</b>  | Zhang, X, <b>40–42, 42, 44,</b><br>46, <b>50, 50, 55–</b><br>58, 73, 75, <b>77,</b><br><b>81, 82, 82, 88,</b><br>96, 98, <b>110, 110,</b><br>126, 128, 132,<br>134, 185, 192                             | Zhou, L, 46, <b>47, 61, 95, 98,</b><br>143  |
| Yan, F, 44, 52, <b>55, 68, 89,</b><br>116, <b>124, 170</b>   | Yuan, C, 50, 108   | Zhang, Y, <b>40, 41, 43, 45, 48,</b><br>55, <b>58, 67, 68,</b><br><b>69, 70, 71, 74,</b>   | Zhou, M, 61, 69, 146, 172   |
| Yan, G, <b>54, 61, 122, 144</b>                              | Yuan, H, <b>50, 52, 108, 115</b>   | Zhang, M, 52, 115  | Zhou, N, <b>61, 145</b>   |
| Yan, H, <b>65, 158</b>                                       | Yuan, Q, 50, 109   | Zhang, N, <b>65, 158</b>   | Zhou, Q, 40, 50, <b>59, 76, 109,</b><br>139   |
| Yan, P, 46, 97   | Yuan, X, <b>40, 77</b>   | Zhang, P, 54, 121  | Zhou, S, 49, 106  |
| Yan, T, 64, 73, 154, 186                                     | Yuan, Y, 42, 52, 55, 58,<br><b>68, 85, 116, 126,</b><br>136, <b>170</b>  | Zhang, Q, 42, 46, <b>57, 59, 68,</b><br>69, <b>74, 83, 95,</b><br><b>132, 137, 169,</b><br>174, <b>189</b>   | Zhou, W, <b>67, 73, 165, 188</b>  |
| Yan, X, <b>62, 66, 148, 162</b>                              | Yuan, Z, 50, 108   | Zhang, R, <b>49, 57, 69, 73,</b><br><b>105, 133, 174,</b><br>187   |   |
| Yan, Y, <b>48, 67, 103, 167</b>                              | Yue, F, 52, 114  | Zhang, S, 46, 50, <b>53, 56,</b><br>67, 96, 109, <b>117,</b><br><b>130, 166</b>  |   |
| Yang, A, 69, 171   | Yue, R, <b>43, 43, 86, 86</b>  | Zhang, T, <b>60, 61, 140, 144</b>  |   |
| Yang, B, <b>50, 58, 109, 135</b>                             | Yue, Y, 48, 103  | Zhang, W, 57, 58, 66, <b>74,</b><br>130, 136, 161,<br><b>191</b>   |   |
| Yang, C, 43, 61, 87, 146                                     | Yuen, KC, 69, 172  | Zhang, X, <b>40–42, 42, 44,</b><br>46, <b>50, 50, 55–</b><br>58, 73, 75, <b>77,</b><br><b>81, 82, 82, 88,</b><br>96, 98, <b>110, 110,</b><br>126, 128, 132,<br>134, 185, 192                             |   |
| Yang, F, <b>63, 152</b>                                      | Yukang, J, 53, 118   | Zhang, Y, <b>40, 41, 43, 45, 48,</b><br>55, <b>58, 67, 68,</b><br><b>69, 70, 71, 74,</b>   |   |
| Yang, G, <b>54, 54, 122, 122</b>                             | Yunlu, J, <b>53, 118</b>   | Zhang, M, 52, 115  |   |
| Yang, H, 41, <b>46, 52, 73, 82,</b><br><b>97, 114, 186</b>   | Zang, Y, 55, <b>68, 124, 170</b>   | Zhang, N, <b>65, 158</b>   |   |
| Yang, J, <b>55, 55, 71, 73, 125,</b><br>125, <b>179, 185</b> |  | Zhang, P, 54, 121  |   |
| Yang, K, <b>46, 47, 97, 100</b>                              |  | Zhang, Q, 42, 46, <b>57, 59, 68,</b><br>69, <b>74, 83, 95,</b><br><b>132, 137, 169,</b><br>174, <b>189</b>   |   |
| Yang, L, 42, <b>72, 85, 185</b>                              |  | Zhang, R, <b>49, 57, 69, 73,</b><br><b>105, 133, 174,</b><br>187   |   |
| Yang, M, 49, 106   |  | Zhang, S, 46, 50, <b>53, 56,</b><br>67, 96, 109, <b>117,</b><br><b>130, 166</b>  |   |
| Yang, P, 68, 170   |  | Zhang, T, <b>60, 61, 140, 144</b>  |   |
| Yang, Q, <b>62, 150</b>                                      |  | Zhang, W, 57, 58, 66, <b>74,</b><br>130, 136, 161,<br><b>191</b>   |   |

- Zhou, X, 40, **42**, 48, 51, **62**,  
66, 71, 76, **83**,  
101, 113, **148**,  
161, 181
- Zhou, Y, 40, **41**, 41, 43, 56,  
58, 63, 67, **72**,  
72, 73, 77, 79,  
**80**, 86, 130, 136,  
151, 152, 165,  
167, **182**, 185,  
186
- Zhou, Z, **43**, **86**
- Zhu, F, **73**, **185**
- Zhu, H, **43**, 56, 60, 61,  
74, **87**, 129, 143,  
144, 189
- Zhu, J, 40, **44**, 44, **45**, 74,  
78, **91**, 91, 189
- Zhu, K, 47, 100
- Zhu, L, 44, 52, 64, 66, 67,  
75, 89, 157, 164,  
165, 192
- Zhu, Q, **67**, **164**
- Zhu, R, **55**, 73, **126**, 188
- Zhu, T, 47, 69, **72**, 100, 171,  
**183**
- Zhu, X, **42**, **50**, 51, **52**, 53,  
60, 62, 64, **66**,  
71, **84**, **110**, 112,  
118, 142, 150,  
157, **163**, 178
- Zhu, Y, 41, **69**, 79, **173**
- Zhu, Z, 40, **65**, 78, **159**
- Zhuang, X, 64, 155
- Zilinskas, R, 55, 124
- Zitnik, M, 50, 108
- Zou, C, 65, 67, 71, 159, 166,  
178
- Zou, H, 47, 98
- Zou, J, 54, 123
- Zou, T, 64, 156

[www.icsa.org](http://www.icsa.org)



International Chinese Statistical Association  
**泛華統計協會**