2018 ICSA
APPLIED STATISTICS
SYMPOSIUM

New Brunswick, NJ
June 14-17, 2018
International Chinese Statistical Association

Applied Statistics Symposium

2018

CONFERENCE INFORMATION, PROGRAM AND ABSTRACTS

June 14 - 17, 2018
Hyatt Regency New Brunswick
New Brunswick, NJ, USA

Organized by
International Chinese Statistical Association

©2018
International Chinese Statistical Association
# Contents

Welcome .................................................................................................................. 1
Conference Information .............................................................................................. 2
  ICSA Officers and Committees .................................................................................. 2
  Conference Committees ............................................................................................ 5
  Acknowledgements ................................................................................................... 7
  Transportation and Parking ......................................................................................... 8
  Map ........................................................................................................................... 9
  Conference Venue Information ................................................................................. 10
  Social Events ........................................................................................................... 12
  Program Overview .................................................................................................... 13
  Keynote Lectures ...................................................................................................... 15
  Banquet Speaker ....................................................................................................... 19
  Student Paper Awards ............................................................................................... 20
  Short Courses ........................................................................................................... 21
  Career Services ......................................................................................................... 26
 Announcements ........................................................................................................... 36
  ICSA 2019 Applied Statistics Symposium ................................................................. 36
  ICSA China 2019 ...................................................................................................... 37
 Scientific Program .................................................................................................... 39
  June 14 2:00 PM - 3:00 PM ....................................................................................... 39
  June 14 3:20 PM - 5:00 PM ....................................................................................... 39
  June 15 8:30 AM - 9:45 AM ..................................................................................... 41
  June 15 10:00 AM - 11:30 AM ................................................................................ 41
  June 15 1:00 PM - 2:00 PM ...................................................................................... 43
  June 15 2:10 PM - 3:50 PM ...................................................................................... 43
  June 15 4:10 PM - 5:50 PM ...................................................................................... 46
  June 16 8:30 AM - 9:30 AM .................................................................................... 48
  June 16 9:50 AM - 11:30 AM .................................................................................. 48
  June 16 1:00 PM - 2:40 PM ...................................................................................... 51
  June 16 3:00 PM - 4:40 PM ....................................................................................... 53
  June 17 8:30 AM - 10:10 AM .................................................................................. 56
  June 17 10:30 AM - 12:10 PM .................................................................................. 59
 Abstracts ....................................................................................................................... 62
  Keynote Lecture I ...................................................................................................... 62
  Session 1: Network Data Analysis ........................................................................... 62
  Session 2: Future of Statisticians in the pharmaceutical world – A panel Discussion with Leaders ........................................................................................................ 62
  Session 3: Some recent development in large-scale statistical learning with complex data ........................................................................................................ 62
  Session 4: SIBS Special Session: Methods for Big Data Analysis in Genomics and Biomedicine ...................................................................................... 63
  Session 5: Basket Trials, Umbrella Trials, Platform Trials and Other Master Protocols ............................................................................................................ 64
  Session 6: Recent Work on Clinical Trial Design and Analysis ................................ 65
  Session 7: Statistical inferences in some semi-parametric models ......................... 66
  Session 8: Recent Advances in Dose Response Clinical Trials ............................... 66
  Session 9: Statistics and machine learning at the frontier of data-driven application and technology ...................................................................................... 67
  Session 10: Data-Driven Decision Making and Objective Statistical Modeling with Big Data Applications ................................................................. 67
  Session 11: Dynamic Modeling and Machine Learning for Big Data from Large Cohort Studies .................................................................................... 68
| Session 12: Deep learning and its mechanism | 68 |
| Session 13: Making sense of data for business: the role of statistics | 69 |
| Session 14: Challenges and Innovations in Modeling Large-Scale Imaging Data | 70 |
| Session 15: Statistical advance in functional genomics | 71 |
| Session 16: Innovative Statistical Designs in Early Oncology Drug Development | 72 |
| Session 17: Recent Advances in Nonparametric Methods | 73 |
| Session 18: Advancing Complex Innovative Clinical Trial Design | 74 |
| Session 19: Real-World Evidence: Moving from Data to Knowledge and Decision Making | 74 |
| Session 20: The role of real world evidence in the development of medical treatment | 75 |
| Session 21: Statistical methods and applications in mobile health | 75 |
| Session 22: Recent Developments in Risk Prediction Models | 76 |
| Session 23: Classification for high dimensional complex data | 76 |
| Session 25: New methods and challenges for analyzing of process data | 77 |
| Session 26: Biosimilar and bioequivalence | 78 |
| Session 27: Clinical trial design and statistical methods for cancer studies | 78 |
| Session 28: Advance in Statistical Methods for Large and Complex Data | 79 |
| Session 29: Statistical Issues for Prognostic Biomarker Discovery and Risk Prediction in Cancer Genomics | 80 |
| Session 30: Revisiting challenges in time-to-event type of analysis in oncology trials | 81 |
| Session 31: Effective Communication | 82 |
| Session 32: Statistical Process Control | 83 |
| Session 33: Decision making in early clinical development of new drugs | 83 |
| Session 34: Statistical inference in data science | 84 |
| Session 35: Advances in the pediatric drug development with extrapolation | 84 |
| Session 36: Model Informed Clinical Drug Development | 85 |
| Session 37: Novel Applications of Statistical Learning Methods in Complex Biomedical Data | 86 |
| Session 38: Statistical Models for Understanding HIV/AIDS Epidemics | 86 |
| Session 39: Logical Inference on Efficacy in Subgroups and Their Mixtures: Targeted Therapies | 87 |
| Session 40: The Jiann-Ping Hsu Invited Session on Biostatistical and Regulatory Sciences | 87 |
| Session 41: Statistical Inference for Distributed Data and Online Streaming Data | 88 |
| Session 42: New developments in microbiome sequencing data modeling and analysis | 89 |
| Session 43: Flexible modeling and inference for complex data | 89 |
| Session 44: Statistical solutions to practical issues in educational testing | 90 |
| Session 45: New Topics in Statistical Learning | 91 |
| Session 46: Futility analysis and interim decision making | 92 |
| Session 47: Recent Biomarker Developments with Applications in Brain Research and Disorders | 93 |
| Session 48: Collaboration Space between Biostatistics and Pharmacometrics | 93 |
| Session 49: Study Designs for Phase I Oncology Dose-Finding Trials | 94 |
| Session 50: Design and execution of late stage oncology trials in modern cancer drug development | 95 |
| Session 51: Challenges and advances in imaging-genetics | 96 |
| Session 52: Integrative analysis of multi-view data with applications to precision medicine | 96 |
| Session 53: Real world evidence in the US health care system | 97 |
| Session 54: Statistical Inference for Discrete and Categorial Data | 98 |
| Session 55: Modern Statistical Development for Biomedical Big Data | 99 |
| Session 56: New advanced methods in functional data analysis | 100 |
| Session 57: Dimension reduction, variable selection and their applications | 100 |
| Session 58: Analysis and computation for complex models | 101 |
| Session 59: Hierarchical generalized linear models in practice | 102 |
| Session 60: Building Successful Sponsor/CRO Relationship - A Biostatistical Perspective | 102 |
| Session 61: Statistical and Regulatory Issues in Human Drug Abuse Studies | 103 |
| Session 62: Statistical Innovations in Big Data Analysis | 104 |
| Session 63: Student Session | 105 |
| Session 64: An introduction to BFF: new statistical inference tools for data science | 106 |

The 2018 ICSA Applied Statistics Symposium is held from Thursday June 14 to Sunday June 17, 2018 at Hyatt Regency New Brunswick, New Jersey, USA. This is the 27th annual symposium for ICSA. The theme of this conference is The New Era of Data Science and Inference, in recognition of a new era for statisticians with different challenges and opportunities. The organizing committees and the ICSA welcome you to the symposium!

The organizing committees have been working diligently to put together a strong and comprehensive program to provide a wealth of activities and opportunities for discussion and exchange. The symposium program contains six short courses and 120 scientific sessions, including four keynote lectures, six featured sessions and two student award paper sessions, as well as exciting social events. We are delighted to have keynote lectures from distinguished statisticians from both government and academia: Dr. Lisa LaVange (University of North Carolina at Chapel Hill), Dr. Dionne Price (US Food and Drug Administration), Dr. Xuming He (University of Michigan), and Dr. Cun-Hui Zhang (Rutgers, The State University of New Jersey). The symposium highlights contributions of statistical, mathematical and computer sciences and their applications to all other disciplines. It brings together the statistical community and scientists from related fields to present, discuss and disseminate research and best practice in statistics and its applications.

With your full support, this symposium attracts more than 600 statisticians working in academia, government, and industry from all over the world. We hope the symposium offers you great opportunities for learning, networking and recruiting, and that you will receive inspirations from the presented research ideas and develop new ones. Social events in this 2018 ICSA Applied Statistics Symposium include the opening mixer (Thursday, June 14 evening) and banquet (Friday, June 15 – banquet speaker is a renowned scholar and statistics historian Dr. Glenn Shafer). We believe this conference will be a memorable, interesting and enjoyable experience for all of you.

The city of New Brunswick, New Jersey is accessible from most cities across the North East America including New York City and Philadelphia. Downtown New Brunswick provides numerous opportunities for dining, shopping and lodging, etc. The oldest campus of Rutgers University in the city features the botanic Rutgers Garden, Zimmerli Art Museum, and Kirkpatrick Chapel. If you want to enjoy relaxing activities during your stay, the State Theatre in the downtown area offers many shows and plays. Finally, the Hyatt Hotel is only two blocks away from the train station, where you can easily travel to NYC or Princeton. It is our sincere hope you take the opportunity to experience these wonderful activities during your stay in New Brunswick.

Thank you for coming to the 2018 ICSA Applied Statistics Symposium in New Brunswick!

Min-ge Xie, on behalf of
2018 Applied Statistics Symposium Executive and Organizing Committees
ICSA Officers and Committees

ICSA 2018 EXECUTIVES AND MEMBERS OF THE COMMITTEES

EXECUTIVES
President: Aiyi Liu (aiyiliu65@gmail.com)
Past President: T. Tony Cai (tcai@wharton.upenn.edu)
President-elect: Heping Zhang (heping.zhang@yale.edu)
Executive Director: Gang Li (2017-2019, gli@its.jnj.com)
ICSA Treasurer: Hongliang Shi (2016-2018, hongliangshi15@gmail.com)

Office of ICSA: Weiliang Qiu, Email: oicsa@icsa.org, Phone: (912) 478-1277.

BOARD of DIRECTORS
Alan Chiang (2016-2018, chiangay@lilly.com),
Yanyuan Ma (2016-2018, yanyuanma@stat.sc.edu),
Lu Tian (2016-2018, lutian@stanford.edu),
Bo Yang (2016-2018, byang2233@gmail.com),
Hao Helen Zhang (2016-2018, hzhang@math.arizona.edu),
Jie Chen (2017-2019, jiechen0713@gmail.com),
Jianhua Hu (2017-2019, jh3992@cumc.columbia.edu),
Bo Huang (2017-2019, bo.huang@pfizer.com),
Chung-Yu Huang (2017-2019, cyhuang@jhu.edu),
Sijian Wang (2017-2019, swang@biostat.wisc.edu),
Chung-Chou H (Joyce) Chang (2018-2020, changj@pitt.edu),
Haitao Chu (2018-2020, chux0051@umn.edu),
Daniel Li (2018-2020, dli.lihe@gmail.com),
Mark Chunming Li (2018-2020, Chunming.M.Li@pfizer.com),
Niansheng Tang (2018-2020, nstang@yru.edu.cn).

Biometrics Section Representative
Lei Liu (2018, lei.liu@northwestern.edu),

STANDING COMMITTEES

Program Committee
Tony Jianguo Sun (Chair, 2018, sunj@missouri.edu), Guoqing Diao (2018-2020, gdiao@gmu.edu), Kai Yu (2018-2020, yuka@mail.nih.gov), Liuquan Sun (2018-2020, slq@amt.ac.cn), Yi Li (2016-2018, yili@umich.edu), Lanju Zhang (2016-2018, Lanju.zhang@AbbVie.com), Hongmei Jiang (2016-2018, hongmei@northwestern.edu), Wenguang Sun (2017-2019, USC, wenguans@marshall.usc.edu), Dan Yang (dyang@stat.rutgers.edu), Hong Tian (2017-2019, HTian@its.jnj.com); Menglin Liu (2018-2020, Mengling.Liu@nyumc.org), Xiwu Lin (2018-2020, xlin38@ITS.JNJ.com), Wenbin Lu (2018-2020, wlu4@ncsu.edu).
ICSA Officers and Committees

Finance Committee
Hongliang Shi (Chair, 2016-2018, hongliangshi15@gmail.com), Linda Yau (2016-2018, yau.linda@gene.com), Rochelle Fu (2019-2021, fur@ohsu.edu).

Nomination for Election Committee
Colin O. Wu (Chair, 2018, NHLBI/NIH, wuc@nhlbi.nih.gov); Joanna Shih (2016-2018, jshih@mail.nih.gov), Tsai-Hung Fan (2016-2018, thfanncu@gmail.com), Hongzhe Li (2017-2019, hongzhe@mail.med.upenn.edu), Niansheng Tang (2017-2019, nstang@ynu.edu.cn); Chin-Tsang Chiang (2018-2020, chiangct@math.ntu.edu.tw); Jeen Liu (2018-2020, Liu_jeen@allergan.com), Xin Tian (2018-2020, tianx@nhlbi.nih.gov), Ming Tan (2018-2020, Ming.Tan@georgetown.edu).

Publication Committee
Yichuan Zhao (2018, yichuan@gsu.edu), Yi Huang (Editor of Bulletin, yihuang@umbc.edu), Mei-Cheng Wang (Co-Editor of SIB, mewang@jhu.edu), Hongzhe Li (Co-Editor of SIB, hongzhe@mail.med.upenn.edu), Hsin-Cheng Huang (Co-Editor of S. Sinica), Ruey Tsay (Co-Editor of S. Sinica), Zhiliang Ying (Co-Editor of S. Sinica), Jiahua Chen (Editor of ICSA book series), Din Chen (Co-Editor of ICSA book series).

Membership Committee
Bo Fu (Chair, 2018, bo.fu@abbvie.com), Jiajuan Liang (2016-2018, JLiang@newhaven.edu), Samuel Wu (2016-2018, sw45@ufl.edu), Pengsheng Ji (2017-2019, psji@uga.edu), Yong Chen (2017-2019, ychen123@upenn.edu), Rui Feng (2017-2019, rui teng@mail.med.upenn.edu), Mark Li (2018-2020, Chunming_M.Li@pfizer.com), Yaping Wang (2018-2020, yapping.wang@fda.hhs.gov), Chenguang Wang (2018-2020, cwang68@jhmi.edu).

Special Lecture Committee
Gang Li of UCLA, (2018, Chair, vli@g.ucla.edu), Qiman Shao (Chair, CUHK, qmshao@sta.cuhk.edu.hk), Yazhen Wang (UW-Madison, yzwang@stat.wisc.edu), Yin Xia (UNC/Fudan, xiayin@email.unc.edu), Ming Yuan (2018-2020, ming.mingyuan@gmail.com), Zhezhen Jin (2018-2020, zj7@cumc.columbia.edu).

Awards Committee
Mei-Cheng Wang (Chair, 2018, Johns Hopkins University, mcwang@jhu.edu), Ming Yen Cheng (2016-2018, cheng@math.ntu.edu.tw), Tsung-Chi Cheng (2016-2018, cheng@nccu.edu.tw), Yazhen Wang (2017-2019, yzwang@stat.wisc.edu), Zongming Ma (2017-2019, zongming@wharton.upenn.edu), Yichuan Zhao (2017-2019, yichuan@gsu.edu), Song-Xi Chen (2017-2019, songchen@iastate.edu), Harrison Zhou (2017-2019, huibin.zhou@yale.edu), Frank Guanghai Liu (2018-2020, guanghan_frank_liu@merck.com), Song Yang (2018-2020, yangso@nhlbi.nih.gov).

IT Committee
Chengsheng Jiang (Chair, 2018, chengsheng.jiang@gmail.com), Jessica Jin (jin_si@hotmail.com), Don Sun (dsun48@gmail.com), Hongtu Zhu (htzhu@email.unc.edu).
ICSA Officers and Committees

Archive Committee
Lili Yu (Chair, 2018, lyu@georgiasouthern.edu), Smiley Cheng, Shein-Chung Chow, Nancy Lo

Lingzi Lu Award Committee (ASA/ICSA)
Jichun Xie (2017-2019, Duke University, jichun.xie@duke.edu), Hao Chen (2017-2019, UC Davis, hxchen@ucdavis.edu), Haonan Wang (2016-2018, wanghn@stat.colostate.edu), Victoria Sides (2017-2019, victoriasides16@gmail.com), Shelly Hurwitz (2017-2019, hurwitz@hms.harvard.edu), Eric Kolaczyk (2014-2017, kolaczyk@math.bu.edu).

ICSA Representative to JSM Program Committee
Wenguang Sun (2018, wenguans@marshall.usc.edu), Zhen Chen (2019, chenzhe@mail.nih.gov).

ICSA China Liaison
Zhezhen Jin (2018-2020, zj7@cumc.columbia.edu)

AD HOC COMMITTEES

2018 Applied Statistics Symposium Committee
Minge Xie, Chair of the Executive Committee (mxie@stat.rutgers.edu).

2018 ICSA China Conference Committee
Gang Li of UCLA, Chair of the Program Committee (vli@g.ucla.edu).

2018 JSM Local Committee
Jiahua Chen (Chair, jhchen@stat.ubc.ca)

Investment Ad Hoc Committee
Linda Yau (Chair, 2018, yau.linda@gene.com), Fang Chen (FangK.Chen@sas.com), Hongliang Shi (hongliangshi15@gmail.com), Ming-Dauh Wang (wang_ming-dauh@lilly.com), Rochelle Fu (fur@ohsu.edu), Mingxiu Hu (Consultant, Mingxiu.Hu@takeda.com)

CHAPTERS

ICSA-Canada Chapter
Li Quan Wang (Chair, LiQuan.Wang@umanitoba.ca)

ICSA-New England Chapter
Ouhong Wang (Chair, Ouhong.Wang@vrtx.com)

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

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

- Min-ge Xie (Rutgers, The State University of New Jersey), Chair
- Ying Hung (Rutgers, The State University of New Jersey), Co-Chair
- Gang Li (Janssen Research and Development), ICSA Executive Director

Program Committee

- Cheng Yong Tang (Temple University), Chair
- Aiyou Chen (Google, LLC.)
- Freda Cooner (Sanofi)
- Yang Feng (Columbia University)
- Zijian Guo (Rutgers, The State University of New Jersey)
- Yanzhi Hsu (Eli Lilly and Company)
- Robert Krafty (University of Pittsburgh)
- Xiwu Lin (Janssen Research and Development)
- Xiaodong Luo (Sanofi)
- Zongming Ma (University of Pennsylvania)
- Zhao Ren (University of Pittsburgh)
- Xiaohua Sheng (FMD K&L)
- Qi Tang (Sanofi)
- Yichao Wu (University of Illinois at Chicago)
- Dan Yang (Rutgers, The State University of New Jersey)
- Han Xiao (Rutgers, The State University of New Jersey)
- Ching-Ray Yu (Pfizer)
- Zhigeng Zhao (Temple University)
- Ping-Shou Zhong (Michigan State University)
- Wenxuan Zhong (University of Georgia)
- Ming Zhou (Bristol-Myers Squibb)
- Ming Zhu (Sanofi)

Registration Committee

- Xiaodong Luo (Sanofi), Chair
- Chengsheng Jiang (University of Maryland, College Park)
- Dan Yang (Rutgers, The State University of New Jersey)
- Zhigeng Zhao (Temple University)
- Ming Zhu (Sanofi)

Short Course Committee

- Jun Yan (University of Connecticut), Chair
- Han Xiao (Rutgers, The State University of New Jersey)
- Ching-Ray Yu (Pfizer)
Conference Committees

Student Paper Award Committee

- Wenxuan Zhong (University of Georgia), Chair
- Ying Hung (Rutgers, The State University of New Jersey), Co-Chair
- Ping Li (Pfizer)
- Zhaoling Meng (Sanofi)
- Xiao Song (University of Georgia)
- Yao Xie (Georgia Institute of Technology)
- Peng Zeng (Auburn University)

Fundraising Committee

- Jason Liao (Merck), Chair
- Ping Li (Pfizer)
- Cindy Lu (Bayer)
- Zhaoling Meng (Sanofi)
- Xiaohua Sheng (FMD K&L)
- Rui Tang (Shire)

Logistic Committee

- Freda Cooner (Sanofi), Chair
- Xiwu Lin (Janssen Research and Development)
- Ping Wang (InsMed)
- Ching-Ray Yu (Pfizer)

Communication Committee

- Han Xiao (Rutgers, The State University of New Jersey), Chair
- Yi Chen (Rutgers, The State University of New Jersey)
- Zijian Guo (Rutgers, The State University of New Jersey)

Treasurer

- Hong Tian (Janssen Research and Development)
- Dan Yang (Rutgers, The State University of New Jersey)

Website

- Ming Zhu (Sanofi), Chair,
- Chengsheng Jiang (University of Maryland, College Park)
The 2018 ICSA Applied Statistics Symposium Program Committees gratefully acknowledge the generous support of our sponsors.

Gold Sponsors

Silver Sponsors

General Sponsors
Transportation and Parking

Local Time

New Jersey is in Eastern Time Zone.

Location

The 2018 Applied Statistics Symposium by the International Chinese Statistical Association (ICSA) will be held at **Hyatt Regency New Brunswick**. The address is Two Albany Street, New Brunswick, New Jersey, USA, 08901.

Airports

The closest airport is **Newark Liberty International Airport**. Two alternative airports in New York City are **LaGuardia Airport** and **JFK - John F. Kennedy International Airport**.

Public Transportation to the Hyatt Regency

- **Public transportation from Newark airport to the hotel:**

  From the airport terminal, follow signs for AIRTRAIN to get to NJ TRANSIT train station. Take Northeast Corridor service from the Newark airport station to New Brunswick train station. The Hyatt hotel is 5 mins' walk from the New Brunswick Train station. The train schedule can be found at [http://www.njtransit.com](http://www.njtransit.com).

- **Public transportation from LaGuardia or JFK to the Hotel takes at least two hours and is not recommended.**

Parking and Driving

- The free parking lots are #26 and #30 (see the map on the next page) behind the College Avenue student center, where the banquet will be held. The Hyatt Regency is located at the intersection of Albany and Neilson ST.

- For more information on driving to campus, please refer to the Rutgers website:

### Opening Mixer:

Thursday, June 14, 6:00 PM – 9:00 PM; Address: Hyatt Regency New Brunswick, Atrium Prefunction
Cost: Free

### Banquet:

Friday June 15, 6:45-9:00 PM; Address: Rutgers Student Center, 126 College Ave, New Brunswick, NJ 08901
Cost: Nonstudent attendee $65; student attendee $30; Guest: $60 (16 or older) or $30 (6 - 15 years old)

**Banquet speaker:** Glenn Shafer, PhD,
University Professor and Board of Governors Professor,
Rutgers, The State University of New Jersey

**Banquet Host:** Richard (Ruey) C. Hwang, Ph.D., M.B.A.,
Executive Director, Clinical Supply Operations, Pfizer Inc.
# Program Overview

## Thursday June 14, 2018

<table>
<thead>
<tr>
<th>Time</th>
<th>Room</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00am-5:00pm</td>
<td>Lobby</td>
<td>Registration</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Short course</strong></td>
</tr>
<tr>
<td></td>
<td>Brunswick C</td>
<td>Full-Day: Preparing Statistician to be Successful Big Data Scientist</td>
</tr>
<tr>
<td></td>
<td>Brunswick D</td>
<td>Full-Day: Analyzing Longitudinal Clinical Trial Data</td>
</tr>
<tr>
<td>8:00am-12:00pm</td>
<td>Garden State A</td>
<td>Half-Day: Statistical Topics in Health Economics and Outcomes Research: Patient-Reported Outcomes, Meta-Analysis, and Health Economics</td>
</tr>
<tr>
<td></td>
<td>Conference B</td>
<td>Half-Day: Artificial Intelligence, Machine Learning, and Precision Medicine</td>
</tr>
<tr>
<td></td>
<td>Conference C</td>
<td>Half-Day: Futility Analyses in Confirmatory Clinical Trials – Methods and Procedures</td>
</tr>
<tr>
<td></td>
<td>Garden State B</td>
<td>Half-Day: Statistical Analysis with Noisy Data - Dealing with Measurement Error and Missing Values</td>
</tr>
<tr>
<td>2:00pm-3:00pm</td>
<td>Regency DEF</td>
<td><strong>Keynote I:</strong> Lisa LaVange, UNC Chapel Hill</td>
</tr>
<tr>
<td>3:00pm-3:20pm</td>
<td></td>
<td><em>break</em></td>
</tr>
<tr>
<td>3:20pm-5:00pm</td>
<td></td>
<td><strong>Parallel Sessions</strong></td>
</tr>
<tr>
<td>6:00pm-9:00pm</td>
<td></td>
<td><strong>Opening Mixer</strong></td>
</tr>
<tr>
<td>6:30pm-9:00pm</td>
<td></td>
<td><strong>ICSA Board meeting</strong> (by invitation only)</td>
</tr>
</tbody>
</table>

## Friday June 15, 2018

<table>
<thead>
<tr>
<th>Time</th>
<th>Room</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:30am-5:00pm</td>
<td>Lobby</td>
<td>Registration</td>
</tr>
<tr>
<td>8:45am-9:45am</td>
<td>Regency DEF</td>
<td><strong>Keynote II:</strong> Dionne Price, FDA</td>
</tr>
<tr>
<td>9:45am-10:00am</td>
<td></td>
<td><em>break</em></td>
</tr>
<tr>
<td>10:00am-11:40am</td>
<td></td>
<td><strong>Parallel Sessions</strong></td>
</tr>
<tr>
<td>11:40am-1:00pm</td>
<td></td>
<td><em>Lunch on own</em></td>
</tr>
<tr>
<td>1:00pm-2:00pm</td>
<td>Regency DEF</td>
<td><strong>Keynote III:</strong> Xuming He, Univ. of Michigan</td>
</tr>
<tr>
<td>2:00pm-2:10pm</td>
<td></td>
<td><em>break</em></td>
</tr>
<tr>
<td>2:10pm-3:50pm</td>
<td></td>
<td><strong>Parallel Sessions</strong></td>
</tr>
<tr>
<td>3:50pm-4:00pm</td>
<td></td>
<td><em>break</em></td>
</tr>
<tr>
<td>4:00pm-5:40pm</td>
<td></td>
<td><strong>Parallel Sessions</strong></td>
</tr>
<tr>
<td>6:45pm-9:00pm</td>
<td></td>
<td><strong>Banquet</strong> (Banquet speaker: Glenn Shafer, Rutgers)</td>
</tr>
</tbody>
</table>

*see program*
Saturday June 16, 2018

8:00am-5:00pm Lobby Registration
8:30am-9:30am Regency DEF Keynote IV: Cun-hui Zhang, Rutgers break
9:30am-9:50am see program Parallel Sessions
9:50am-11:30am see program Parallel Sessions
11:30am-1:00pm see program Parallel Sessions
1:00pm-2:40pm see program Parallel Sessions
2:40pm-3:00pm
3:00pm-4:40pm

Sunday June 17, 2018

8:00am-10:30am Lobby Registration
8:30am-10:10am see program Parallel Sessions break
10:10am-10:30am see program Parallel Sessions
10:30am-12:10pm see program Parallel Sessions
Lisa LaVange, Ph.D.
Professor and Associate Chair
Department of Biostatistics in the Gillings School of Global Public Health
The University of North Carolina at Chapel Hill

Dr. Lisa LaVange is Professor and Associate Chair of the Department of Biostatistics in the Gillings School of Global Public Health at the University of North Carolina at Chapel Hill. She is also director of the department’s Collaborative Studies Coordinating Center (CSCC), overseeing faculty, staff, and students involved in large-scale clinical trials and epidemiological studies coordinated by the center. From 2011 to 2017, Dr. LaVange was director of the Office of Biostatistics in the United States Food and Drug Administration (FDA) Center for Drug Evaluation and Research (CDER). There, she oversaw more than 200 statisticians and other staff members involved in the development and application of statistical methodology for drug regulation. She was a leader in developing and assessing the effectiveness and appropriateness of innovative statistical methods intended to accelerate the process from drug discovery to clinical trials to FDA approval and patients’ benefit, with a particular focus on rare diseases. Prior to her government and academic experience, she spent 16 years in non-profit research and 10 years in the pharmaceutical industry. Dr. LaVange is an elected fellow of the American Statistical Association (ASA) and is the 2018 ASA President.

Title: Reflections on Statistical Practice, Leadership, and Training

Location and Time: Regency DEF, Thursday June 14, 2:00pm – 3:00pm.

Abstract:
Having recently completed six years in a statistical leadership role at FDA, I will reflect on the importance and impact of statistics in drug development and regulation, citing examples from policy development and dissemination in addition to statistical reviews of new drugs. Opportunities for FDA statisticians to learn about and practice leadership in their work environment will be compared and contrasted with opportunities in other employment sectors. Training statisticians to lead, formally and informally, is essential for the future of our profession, and I will offer my view for ways this can be accomplished, including highlighting the planned initiatives of the ASA Statistical Leadership Institute.
Dr. Dionne Price received her MS in Biostatistics from the University of North Carolina at Chapel Hill and a PhD in Biostatistics from Emory University. Dr. Price is an active member of the American Statistical Association having served as Chair, Program-Chair and Secretary of the Biopharmaceutical Section. She has also served on the Regional Committee and Regional Advisory Board of the Eastern North American Region of the International Biometrics Society.

Dr. Price is the acting Deputy Director of the Office of Biostatistics in the Office of Translational Sciences, Center for Drug Evaluation and Research, FDA. During her FDA career, Dr. Price has played an active role in the development and application of methodology used in the regulation of anti-infective, anti-viral, ophthalmology, transplant, analgesia, anesthesia, and addiction drug products. She has made significant contributions toward advancing the mission of the FDA through her involvement in cross-cutting efforts impacting numerous stakeholders.

Title: The Changing Landscape of Drug Development: To Adapt or Not to Adapt

Location and Time: Regency DEF, Friday June 15, 8:45am – 9:45am.

Abstract:
Statistics may need to adapt to meet the demands of the changing landscape of drug development. This need to adapt has been motivated by numerous factors including the new era of big data as well challenging problems in various therapeutic areas. The Food and Drug Administration is poised to meet the demands of the changing landscape as evident by various commitments outlined in recent legislation such as the 21st Century Cures Act and the Prescription Drug User Fee Act (PDUFA) reauthorization for fiscal years 2018-2022, known as PDUFA VI. These commitments encompass a broad range of topics including real world evidence, complex innovative designs, patient-focused drug development, and model informed drug development. To achieve the common goal of advancing drug development to meet the evolving needs of patients, statistical input, collaboration, and leadership will be germane. This talk will provide an overview of the commitments and discuss how the Office of Biostatistics is adapting to propel drug development forward in this new era.
Xuming He
H. C. Carver Collegiate Professor at the University of Michigan.

Dr. He earned a bachelor's degree in Applied Mathematics from Fudan University in 1984. He went to graduate school at University of Illinois at Urbana-Champaign and received a Master's degree in Mathematics and then Ph.D. in Statistics in 1989. He joined the University of Michigan as H. C. Carver Collegiate Professor in 2011. His prior appointments include faculty positions at National University of Singapore and University of Illinois at Urbana-Champaign. His research interests include theory and methodology in robust statistics, semiparametric models, quantile regression, data depth, dimension reduction, and subgroup analysis. His interdisciplinary research aims to promote the better use of statistics in biosciences, climate studies, dysphagia research, and social-economic studies. Dr. He is Fellow of the American Association for the Advancement of Science, the American Statistical Association, the Institute of Mathematical Statistics, and the International Statistical Institute.

Title: How good is your selected subgroup?

Location and Time: Regency DEF, Friday June 15, 1:00pm – 2:00pm.

Abstract: Subgroup analysis is frequently used to account for the treatment effect heterogeneity in clinical trials. When a treatment is seen marginally effective for the population of the original study, it is tempting to consider post hoc subgroup identification. When a highly promising subgroup is selected this way, serious questions have to be asked about the potential risks and rewards of the subgroup pursuit. In this talk, we will discuss factors that have direct impacts on the credibility of subgroup pursuit, and then propose a model-free approach to quantify how likely the promise of the selected subgroup is a statistical artifact, and how good the selected subgroup really is. The proposed quantitative analysis of subgroup pursuit can help inform better decisions about any selected subgroup in clinical trials. The talk is based on joint work with Xinzhou Guo.
Cun-Hui Zhang

Distinguished Professor at Rutgers University.

Cun-Hui Zhang is Distinguished Professor of Statistics and Biostatistics at Rutgers University. His research interests include high-dimensional data, machine learning, empirical Bayes, bootstrap, data sketch and hashing, nonparametric methods, multivariate analysis, survival data, functional MRI, closed loop diabetes control, and network tomography. He is a Fellow of the Institute of Mathematical Statistics and a Fellow of American Statistical Association. He is Editor of Statistical Science and serves on the editorial boards of Annals of Statistics, Bernoulli, Statistica Sinica, and Statistics Surveys.

Title: Statistical Inference with High-Dimensional Data

Location and Time: Regency DEF, Saturday June 16, 8:30am-9:30am.

Abstract: We consider statistical inference in a semi-low-dimensional approach to the analysis of high-dimensional data. The relationship between this semi-low-dimensional approach and regularized estimation of high-dimensional objects is parallel to the more familiar one between semiparametric analysis and nonparametric estimation. Low-dimensional projection methods are used to correct the bias of regularized high-dimensional estimators, leading to efficient point and interval estimation. Bootstrap can be used to carry out simultaneous inference. Only a small fraction of labelled data are needed in a semisupervised setting. Examples include regression and graphical models for continuous and binary data.
Glenn Shafer, Ph.D.
University Professor and Board of Governors Professor
Rutgers, The State University of New Jersey

Dr. Glenn Shafer has been a university educator for 40 years. He is best known for his work in the 1970s and 1980s on the Dempster-Shafer theory, an alternative theory of probability that has been applied widely in engineering and artificial intelligence. There is now an entire society devoted to the advancement of this theory, the Belief Functions and Applications Society, which has been holding international conferences since 2010. Glenn is also known for his initiation, with Vladimir Vovk, of the game-theoretic framework for probability. They are now preparing a second edition of their book on the topic, *Probability and Finance: It’s Only a Game!* In 2009, Glenn was recognized for his work on these topics with an honorary doctorate in economics by the University of Economics, Prague. He has been a Guggenheim Fellow, a Fulbright Fellow, and a visiting professor in Paris and Berlin. In 2004, he received Rutgers’ most prestigious faculty award, the Gorenstein Award for Research and Service, and he is one of two dozen Rutgers faculty members recognized as Board of Governors Professors. Glenn also served as Dean of the Rutgers Business School – Newark and New Brunswick from January 2011 to December 2014.

**Title:** Have we been here before?
* (When data is a science, who needs probability?)

**Location and Time:** Rutgers Student Center, 126 College Ave, New Brunswick, NJ 08901
6:45pm-9:00pm, Friday June 15, 2018

**Abstract:**
The science of statistics was first created in Germany in the 18th century, inspired by a proliferation of data. Every century since has seen explosive advances in availability of data and technology to handle it. Each advance has brought new hope that data alone, without assessment of uncertainty, can give us the answers we want.
Jiann-Ping Hsu Pharmaceutical and Regulatory Sciences Student Paper Award

Rui Duan, University of Pennsylvania
— Title: A Fast Score Test for Generalized Mixture Models
— Time: June 16th 9:50 AM – 11:30 AM
— Session 63: Student Session 1 (Conference I)

ICSAS Student Paper Awards

Tonghui Yu, Nanyang Technological University
— Title: Conditional Modeling of Survival Data with Semi-Competing Risks
— Time: June 16th 9:50 AM – 11:30 AM
— Session 63: Student Session 1 (Conference I)

Suzanne Thornton, Rutgers University
— Title: Approximate Confidence Distribution Computing: An Effective Likelihood-Free Method with Statistical Guarantees
— Time: June 16th 9:50 AM – 11:30 AM
— Session 63: Student Session 1 (Conference I)

Huichen Zhu, Columbia University
— Title: Generalized Integrative Principal Component Analysis for Multi-Type Data with Block-Wise Missing Structure
— Time: June 16th 9:50 AM – 11:30 AM
— Session 63: Student Session 1 (Conference I)

Xinlian Zhang, University of Georgia
— Title: Bayesian Spline Smoothing with Ambiguous Penalties
— Time: June 16th 13:00 PM – 14:40 AM
— Session 65 (Regency A)

Xiaowu Dai, University of Wisconsin-Madison
— Title: Optimal Calibration for Computer Model Prediction with Finite Samples
— Time: June 16th 13:00 PM – 14:40 AM
— Session 76: Student Session 2 (Conference I)

Rui Xie, University of Georgia
— Title: Online Sequential Leveraging Sampling Method for Streaming Time Series Data
— Time: June 16th 13:00 PM – 14:40 AM
— Session 76: Student Session 2 (Conference I)

Chixiang Chen, Penn State College of Medicine
— Title: Empirical-Likelihood-Based Criteria for Model Selection on Marginal Analysis of Longitudinal Data with Dropout Missingness
— Time: June 16th 13:00 PM – 14:40 AM
— Session 76: Student Session 2 (Conference I)
SC02: Statistical Topics in Health Economics and Outcomes Research: Patient-Reported Outcomes, Meta-Analysis, and Health Economic Analysis

Length: Half-day

Instructors: Dr. Joseph C. Cappelleri (Pfizer Inc); Prof. Thomas Mathew (University of Maryland Baltimore County)

Outline/Description: Based in part on the recently published co-edited volume Statistical Topics in Health Economics and Outcomes Research, this four-hour short course recognizes that, with ever-rising healthcare costs, evidence generation through health economics and outcomes research (HEOR) plays an increasingly important role in decision-making about the allocation of resources. This course highlights three major topics related to HEOR, with objectives to learn about 1) patient-reported outcomes, 2) analysis of aggregate data, and 3) methodological issues in health economic analysis. Key themes on patient-reported outcomes are presented regarding their development and validation: content validity, construct validity, exploratory factor analysis, confirmatory factor analysis, person-item maps, and reliability. Regarding analysis of aggregate data, several areas are elucidated: traditional meta-analysis, network meta-analysis, model validation, meta-regression, and best practices for the conduct and reporting of aggregated data. For methodological issues on health economic analysis, cost-effectiveness criteria are covered: traditional measures of cost-effectiveness, the cost-effectiveness acceptability curve, statistical inference for cost-effectiveness measures, the fiducial approach (or generalized pivotal quantity approach), and a probabilistic measure of cost-effectiveness. Examples are illustrated throughout the course to complement the concepts. Attendees are expected to have at least basic quantitative knowledge.

The learning objectives are to understand and critique the major methodological issues in outcomes research on the development and validation of patient-reported outcomes, traditional meta-analysis and network meta-analysis, and health economic analysis.

References:


About the Instructors: Joseph C. Cappelleri, PhD, MPH, MS, is an executive director in the Statistical Research and Data Science Center at Pfizer Inc. He earned his MS in statistics from the City University of New York, Ph.D. in psychometrics from Cornell University, and M.P.H. in epidemiology from Harvard University. As an adjunct professor, Dr. Cappelleri has served on the faculties of Brown University, University of Connecticut, and Tufts Medical Center. He has delivered numerous conference presentations and has published extensively on clinical and methodological topics (over 1,200 co-authored publications and conference presentations), including on regression-discontinuity designs, meta-analyses, and health measurement scales. He is lead author of the book Patient-Reported Outcomes: Measurement, Implementation and Interpretation, co-author of the monograph Phase II Clinical Development of New Drugs, and a co-editor of the volume Statistical Topics in Health Economics and Outcomes Research. Dr. Cappelleri is a Fellow of the American Statistical Association.

Thomas Mathew, PhD, is Professor, Department of Mathematics & Statistics, University of Maryland Baltimore County (UMBC). He earned his PhD in statistics from the Indian Statistical Institute in 1983, and has been a faculty member at UMBC since 1985. He has delivered numerous conference presentations, nationally and internationally, and has published extensively on methodological and applied topics, including cost-effectiveness analysis, bioequivalence testing, exposure data analysis, meta-analysis, mixed and random effects models, and tolerance intervals. He is the co-author of two books Statistical Tests in Mixed Linear Models and Statistical Tolerance Regions: Theory, Applications and Computation, both published by Wiley. He has served on the Editorial Boards of several journals, and is currently an Associate Editor of Sankhya and the Journal of the American Statistical Association. Dr. Mathew is a Fellow of the American Statistical Association, and a Fellow of the Institute of Mathematical Statistics. He has also been appointed as Presidential Research Professor at his campus.

Target Audience: Statisticians, data scientists, epidemiologists, outcomes researchers, health economists, and healthcare policy and decision-makers.
SC03: Statistical Analysis with Noise Data — Dealing with Measurement Error and Missing Values

**Length:** Half-day  
**Instructor:** Dr. Grace Y. Yi (University of Waterloo)  
**Outline/Description:** Thanks to the advancement of modern technology in acquiring data, data with diverse features and big volume are becoming more accessible than ever. A very important concern on analyzing such data is the quality and provenance of the data. Typically, the challenges caused by measurement imprecision and missing observations are particularly intriguing. Measurement error and missing data arise ubiquitously from various fields including health sciences, epidemiological studies, survey research, economics, and so on. They have been a long standing concern in data analysis and have attracted extensive research interest over the past few decades.

It has been well documented that ignoring measurement error or missing data in statistical analyses may lead to erroneous or even misleading results. The effects of measurement error or missing data are, however, complex, affected by various factors. The objective of this course is to lead the audience to visit these challenging but exciting areas. Specifically, the impact of measurement error and missing data will be demonstrated and different types of measurement error models and missing mechanisms will be discussed. Typical inference strategies for handling measurement error and missing data will be described. The discussion will be illustrated with examples and applications.

The participants of the short course are entitled with a discount of purchasing the following reference book which expands the course material.

**References:**  

**About the Instructor:** Grace Y. Yi is Professor of Statistics and University Research Chair at the University of Waterloo. Grace received her Ph.D. in Statistics from the University of Toronto in 2000. She is a Fellow of the American Statistical Association and an Elected Member of the International Statistical Institute. She is the Editor-in-Chief of The Canadian Journal of Statistics (2016-2018). She was President of the Biostatistics Section of the Statistical Society of Canada in 2016, and the Founder and Chair of the first chapter (Canada Chapter) of the International Chinese Statistical Association. Her research interests focus on developing statistical methodology to address various challenges arising from medical studies, clinical trials, epidemiology and survey research. Her recent research has been particularly centering around problems concerning missing data, measurement error in variables, and high-dimensional complex-structured data. Grace was the 2010 winner of the CRM-SSC Prize, an honor awarded in recognition of a statistical scientist’s professional accomplishments in research during the first 15 years after having received a doctorate. She was a recipient of the prestigious University Faculty Award granted by the Natural Sciences and Engineering Research Council of Canada (NSERC). Her work with Xianming Tan and Runze Li won The Canadian Journal of Statistics Award for 2016. Grace has served as an Associate Editor for statistical journals, including The Canadian Journal of Statistics, The Journal of the Royal Statistical Society (Series C), The Journal of Applied Probability, Statistics in Biosciences, STAT, and Biostatistics and Epidemiology.

**Target Audience:** Researchers interested in this topic, subject-matter analysts and graduate students.

SC04: Futility Analyses in Confirmatory Clinical Trials — Methods and Procedures

**Length:** Half-day  
**Instructors:** Dr. Paul Gallo (Novartis); Dr. Satrajit Roychoudhury (Pfizer Inc)  
**Outline/Description:** Futility analyses (FA) are increasingly utilized in clinical trials. FA involves interim evaluation of the trial’s primary hypothesis to determine if there is a low probability of a positive result with trial continuation, or if the desired clinically meaningful effects can already be ruled out with reasonable confidence. FA can improve resource efficiency by the halting of trials with ineffective interventions and enabling sponsors to redirect efforts to more promising pursuits. FA also have ethical advantages in that fewer trial participants may be exposed to ineffective and possibly toxic interventions, and public health advantages in that trial results may be conveyed to the medical community in a more timely fashion.

FA should be carefully planned during trial design and described in the protocol as there are important statistical and operational consequences. Concerns include the control of statistical error rates and the concern for operational bias resulting from interim evaluations. There are varied and expanding statistical tools available for FA. Challenging questions arise during trial design regarding how FA should be conducted, a threshold at which futility would be established, and when futility should be assessed. Non-constancy of effect size and familiar limitations of accruing interim data can raise further challenges. Data Monitoring Committees play a pivotal role in futility evaluation. Ensuring DMC access to appropriate data, ensuring DMC member understanding of futility methodologies, and thoughtful and efficient DMC reports describing FA are important for optimal recommendations.

In this course, we will describe current practices and recent advances in methodological approaches and procedural issues, and illustrate with examples and case studies. We describe what FA are, why they are conducted, where and when they should be considered, and how they should be methodologically and operationally conducted.

This course is for statistical researchers working in the pharmaceutical industry, academic institutions, or regulatory
Short Courses

SC08: Analyzing Longitudinal Clinical Trial Data

Length: Full-day
Instructors: Dr. Craig Mallinckrodt (Eli Lilly & Company); Dr. Ilya Lipkovich (IQVIA)
Outline/Description: The statistical theory relevant to analyses of longitudinal clinical trial data is extensive, and applying that theory in practice can be challenging. Therefore, this course focuses on the most relevant and current theory, using practical and easy to implement approaches for bringing that theory into routine practice. Emphasis is placed on examples with realistic data and the programming code to implement the analyses is provided, usually in both SAS and R. This course closely follows the content of a recent book by the instructors.

Although this course focuses on analytic methods, analyses cannot be considered in isolation, but rather as part of a holistic approach to study development and implementation. That approach begins with determining objectives, followed by choosing estimands, design, analyses, and assessing sensitivity. The intent of this course is to facilitate an integrated understanding of key concepts from across the study development process — through an example-oriented approach.

Learning objectives of the course include:

1. Be able to apply key theory relevant to understanding the merits of various longitudinal analytic approaches in differing circumstances.
2. Be able to implement broadly useful longitudinal analytic approaches using simple code in commercial software.
3. Understand the three pillars of preventing and treating missing data, with emphasis on a structured approach to study development, and be able to apply the three pillars principles and the structured approach to study development to their own research
4. Understand basic principles of sensitivity analyses for incomplete data, and be able to implement key sensitivity analyses, including reference-based imputation and delta adjustment tipping point approaches.

Part I covers key concepts and considerations applicable to modeling longitudinal data. Part II focuses on study development, including defining objectives and estimands, and preventing missing data, which is an inevitable problem in clinical trials. Part III goes into detail on analytic approaches to account for missing data and for conducting sensitivity analyses. Part IV integrates key ideas from Parts I-III to illustrate a comprehensive approach to study development and analyses of realistic data sets.

Throughout the course, example data sets are used to illustrate and explain key analyses and concepts. These data sets were constructed by selecting patients from actual clinical trial data sets and manipulating the observations in ways useful for illustration. By using small data sets attendees can more easily understand exactly what an analysis does and how it does it. For the comprehensive study development and analysis example in Part IV two data sets contrived from actual clinical trial data are used to further illustrate key points for implementing an overall analytic strategy that includes sensitivity analyses and model checking.

References:

About the Instructors: Dr. Mallinckrodt received his PhD in 1993 from Colorado State University, where he subsequently held a joint appointment in the departments of statistics and clinical sciences. Craig joined Lilly in 1998 and has extensive drug development experience covering all four agencies, or statistics students. The course is oriented towards users (i) who are interested to implement futility analysis in confirmatory trials; (ii) who are interested to learn effective ways to communicate with clinical teams regarding implementation of futility analysis in practice. Upon completing this course, participants will know how to implement futility analysis for confirmatory trials; know how to use available statistical packages for futility analysis; be able to interact and communicate efficiently with other stakeholders.

About the Instructors: Dr. Paul Gallo is a Senior Biometrical Fellow and a member of the Statistical Methodology Group at Novartis Pharmaceuticals in East Hanover, NJ. He received a Ph.D. in Statistics from the University of North Carolina in 1982. A main activity has involved support of Data Monitoring Committees; he has worked closely with DMCs in many Novartis trials, and authored company process documents and technical guidelines governing interim analyses. He has been active in industry initiatives relating to adaptive trials, with particular focus on interim monitoring and DMC process issues. He is a Fellow of the American Statistical Association.

Dr. Satrajit Roychoudhury is a Senior Director and a member of Statistical Research and Innovation group in Pfizer Inc. Prior to joining, he was a member of Statistical Methodology and consulting group in Novartis. He started his career as a research statistician in Schering Plough Research Institute (now Merck Co.). He has 10+ years of extensive experience in working with different phases of clinical trial. His primary expertise includes implementation of innovative statistical methodology in clinical trial. He co-authored several publications/book chapters in this area and provided statistical training in major conferences. His area of research includes the use of model based approaches and Bayesian methods in clinical trials.

Target Audience: Statistical researchers working in the pharmaceutical industry, academic institutions, or regulatory agencies involved in planning and performing futility analysis in clinical trial.
clinical phases in multiple therapeutic areas. Dr. Mallinckrodt has published extensively on missing data. He led the PhRMA expert team on missing data and currently leads the Drug Information Association Scientific Working Group on missing data. Dr. Mallinckrodt is a Fellow of the American Statistical Association and recently won the Royal Statistical Society’s award for excellence in the pharmaceutical industry for his book titled A Practical Guide to the Prevention and Treatment of Missing Data.

Dr Lipkovich is Principal Scientific Advisor at IQVIA. Ilya received his PhD in Applied Statistics from Virginia Polytechnic Institute and State University in 2002. He has more than 15 years of statistical consulting experience in pharmaceutical industry. Ilya’s research interests include clustering, predictive modeling and subgroup identification in clinical data, analysis with missing data, and causal inference from observational data. Ilya is a co-developer of novel subgroup identification methods (SIDES) and chairs the QSPI Subgroup Analysis Working Group sponsored by the Society of Clinical Trials. He is a widely published author and frequent presenter at conferences, and a co-author of the book “Analyzing Longitudinal Clinical Trial Data. A Practical Guide” (CRC Press).

**Target Audience:** Practitioners, students, and researchers interested in longitudinal analytic approaches for clinical trial data.

**SC09: Preparing Statistician to be Successful Big Data Scientist**

**Length:** Full-day  
**Instructors:** Dr. Hui Lin (DowDuPont); Dr. Ming Li (Amazon)

**Outline/Description:** With recent big data revolution, enterprises ranging from FORTUNE 500 to startups across the US are hungry for data scientists to bring valuable business insight from all the data collected. Statisticians are great data scientist candidates, but there are relatively few data scientists with statistics education background. In this short course, we will walk through the needed data science knowledge and skills through hands-on exercises to prepare statisticians to be successful data scientists. Data access and manipulation (i.e., extract-transform-load) in production environments are typical gaps for statisticians, and these topics will also be in this course. Data science is a combination of science and art with data as the foundation. We will also include the art part to guide participants to learn typical data science project flow, general pitfalls in data science projects, and soft skills to communicate with business stakeholders effectively. The Databricks community edition cloud platform and R-Studio will be used to cover programming and platform (such as Spark, Hadoop, GPU, SQL, and R) and machine learning algorithms (including examples for supervised learning, unsupervised learning, and deep learning). The prerequisite knowledge is MS level education in statistics and entry level of R knowledge.

This is an enhanced version of a similar highly-rated and full-day training course (CE 11C) offered at JSM 2017 in Baltimore with updated material to reflect students suggestions and new trends in data science. The following is a list of topics included in the course:

1. **Introduction to Data Science.** In this section, we will introduce the history and trends in data science. We will list common requirements for successful data scientist and evaluate for participants to find the skill gaps and give recommendations to bridge the gaps. Participants will have a good understanding of what data scientists do and know their skill and knowledge gaps after taking this section.

2. **Big Data Cloud Platform.** In this section, we will first introduce how to open Databricks community edition account to leverage its scalable cloud-based big data platform (i.e., a mini Hadoop/spark or GPU cluster with R, Python, SQL and Markdown capability). Nearly all the programming and demo related topics are based on this cloud platform. We will briefly introduce SQL and databases, Hadoop MapReduce, Hive, Spark and GPU computing environment for high-performance computing with big data. Participants will be confident to work with these big data platforms after this section.

3. **Data Science Project Cycle.** In this section, we will leverage the Databricks platform Spark capability with R to go through some critical topics in the data science cycle. The platform provides a production-like environment at enterprises. Participants will be able to use R to leverage Spark to do those data manipulation after this section.

4. **Machine Learning Algorithms.** In this section, we will leverage the Databricks platform to explore machine learning algorithms typically used in industries with hands-on examples. Participants will learn how to apply these algorithms to real-world problems through big data platform after this section.

5. **Soft Skills for Data Scientists.** In this section, we will introduce the needed soft skills that are essential in data science projects in enterprise environments. We will talk about necessary project management skills with agile concepts and how to effectively communicate with business partners to define and solve data science problems. We will illustrate how to lead with confidence given the strong technical background that statisticians have. Participants will be familiar with data science collaborations in enterprise environments after this section.

**About the Instructors:** Dr. Hui Lin is currently Data Scientist at DowDuPont. She is a leader in the company at applying advanced data science to enhance Marketing and Sales Effectiveness. She has been providing statistical leadership for a broad range of predictive analytics and market research analysis since 2013. She is the co-founder of Central Iowa R User Group, blogger of scientistcafe.com and 2018 Program Chair of ASA Statistics in Marketing Section. She enjoys making analytics accessible to a broad audience and teaches tutorials and workshops for practitioners on data science. She holds MS and Ph.D. in statistics from Iowa State University, BS in mathematical statistics from Beijing Normal Unives-
Dr. Ming Li is currently a Senior Data Scientist at Amazon and an Adjunct Faculty of Department of Marketing and Business Analytics in Texas A&M University – Commerce. He is the Chair of Quality & Productivity Section of ASA for 2017. He was a Data Scientist at Walmart and a Statistical Leader at General Electric Global Research Center. He obtained his Ph.D. in Statistics from Iowa State University at 2010. With deep statistics background and a few years’ experience in data science, he has trained and mentored numerous junior data scientist with different background such as statistician, programmer, software developer, database administrator and business analyst. He is also an Instructor of Amazon’s internal Machine Learning University and was one of the key founding member of Walmart’s Analytics Rotational Program which bridges the skill gaps between new hires and productive data scientists.

**Target Audience:**
(1) statistician in traditional industry sectors such as manufacturing, pharmaceutical and banking; (2) statistician in government agencies; (3) statistical researchers in universities; and (4) Ph.D. and M.S. graduate students in statistics and other departments.

---

### SC11: Artificial Intelligence, Machine Learning, and Precision Medicine

**Length:** Half-day

**Instructor:** Dr. Haoda Fu (Eli Lilly & Company)

**Outline/Description:** This half day short course will provide an overview of statistical machine learning and artificial intelligence techniques with applications to precision medicine, in particular to deriving optimal individualized treatment strategies for precision medicine. This short course will cover both treatment selection and treatment transition. The treatment selection framework is based on outcome weighted classification. We will cover logistic regression, support vector machine (SVM), robust SVM, and angle based classifiers for multi-category learning. We will show how to modify these classification methods into outcome weighted learning algorithms for precision medicine. The second part of short course will cover treatment transition. We will provide an introduction on reinforcement learning techniques. Algorithms, including dynamic programming for Markov Decision Process, temporal difference learning, SARSA, Q-Learning algorithms, and actor-critic methods, will be covered. We will discuss on how to use these methods for developing optimal treatment transition strategies. The techniques discussed will be demonstrated in R.

**Target Audience:**Graduate students who have some knowledge of statistics and want to be introduced to statistical machine learning, or practitioners who would like to apply statistical machine learning techniques to their problems on personalized medicine and other biomedical applications.

---

**About the Instructor:** Dr. Haoda Fu is a research advisor and a stats group leader for Machine Learning, Artificial Intelligence, and Digital Connected Care from Eli Lilly and Company. He is also an adjunct professor of biostatistics, Indiana university school of medicine. Dr. Fu received his Ph.D. in statistics from University of Wisconsin – Madison in 2007 and joined Lilly after that. Since he joined Lilly, he has been very active in statistical methodology research. He has more than 70 publications in areas such as Bayesian adaptive design, survival analysis, personalized medicine, indirect and mixed treatment comparison, joint modeling, Bayesian decision making, and drug safety evaluation for rare events. In recent years, his research area focuses on machine learning and artificial intelligence.
2018 ICSA Applied Statistics Symposium Career Service

Below are the ICSA 2018 Symposium gold sponsors which currently have statistics related openings. For ICSA 2018 Symposium attendees, you are encouraged to submit your resumes/CVs to the corresponding email address listed below to expedite the application review process, unless specified otherwise. For position details, please check with the contact person of each company directly.

ICSA Career Service offers interviewing facility for our gold sponsors and career-seekers. Companies will conduct interviews on-site during the Symposium. For interview scheduling, please check with the contact person of each company directly.

<table>
<thead>
<tr>
<th>Company</th>
<th>Contact Name</th>
<th>E-mail Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boehringer Ingelheim</td>
<td><a href="https://careers.boehringer-ingelheim.com">https://careers.boehringer-ingelheim.com</a></td>
<td></td>
</tr>
<tr>
<td>Brightech</td>
<td>Tania Valenzuela</td>
<td><a href="mailto:tvalenzuela@brightech-intl.com">tvalenzuela@brightech-intl.com</a></td>
</tr>
<tr>
<td>FMD K&amp;L</td>
<td>Xiaohua Sheng</td>
<td><a href="mailto:Xiaohua.Sheng@klerv.com">Xiaohua.Sheng@klerv.com</a></td>
</tr>
<tr>
<td>Johnson &amp; Johnson</td>
<td>James Pan</td>
<td><a href="mailto:JPan3@its.jnj.com">JPan3@its.jnj.com</a></td>
</tr>
<tr>
<td>Lotus Group</td>
<td></td>
<td><a href="mailto:info@tlgcareers.com">info@tlgcareers.com</a></td>
</tr>
<tr>
<td>TechData Services</td>
<td>Ju Zhang</td>
<td><a href="mailto:ju.zhang@techdataservice.com">ju.zhang@techdataservice.com</a></td>
</tr>
</tbody>
</table>
At Boehringer Ingelheim our vision "Value through Innovation" shows our innovative, sustainable and successful entrepreneurial approach. To learn more about our pipeline and career opportunities, visit our website at…

https://careers.boehringer-ingelheim.com

**Senior Biostatistician or Principal Biostatistician:**

You will perform duties of a Trial Statistician to support complex clinical trials within national or international development projects or for marketed products as required. Assure well-designed clinical trials. Provide statistical expertise necessary to design, analyze, interpret and communicate the results of complex clinical trials. Provide statistical input for publications on clinical trials. Either support Project Statisticians on complex projects or act as a Project Statistician for early projects, backup projects, or projects with established BI experience. Act as a team leader for a complex project or mega-trial.


**Sr. Principal Biostatistician:**

You will perform duties of a Project or Trial Statistician to support high-profile international clinical development projects/trials or marketed products as required. Provide statistical input into the creation of clinical development plans and/or marketing and publication strategies. Assure well-designed clinical trials. Take on the key statistical responsibility in the planning and preparation of regulatory submissions. Represent the company as required before regulatory agencies and other organizations in defense of statistical methodologies and data analyses related to our products.


**Sr. Principal Methodology Statistician**

You will act as an international specialist in applied statistical methodology to foster methodological innovation in BI Clinical Statistics and develop innovative, efficient statistical methods for ready-to-use application in BI studies and projects. Identify and evaluate innovative statistical methods for efficient clinical trial design and analysis and transform promising methods into ready-to-use toolkits. Promote broad usage of innovative, complex statistical methods and mentor P/TSTATs in applying them to their projects. Provide training, consultancy and perform prototype usage of complex or innovative methods. Lead and/or participate in international BI statistics working groups on standardization and innovation of methodology. Mentor statisticians on the technical-methodological track.

About Us
We are a highly reputable Data Service CRO company, working with over 50+ clients since 2004. We have over 100+ employees in US, China and Taiwan and we are committed to always reach the highest level of client satisfaction.

What We Offer
- Biostatistics (Study Design, FDA Interaction)
- SAS Programming (SDTM/ADaM/TFLs and FDA Submission)
- Data Management
- Integrated eClinical Suite CIMS™
- FSP, Deliverable-Based Services
Come join us for the following Speaker Session #60:

**Building A Successful Sponsor/CRO Relationship – A Biostatistical Perspective**

June 16th at 9:50-11:30 AM

Stop by our booth to learn about open positions within our company

We need individuals with skills and motivation to consistently and adaptively lead, connect, shape, and deliver. There will be endless opportunities to successfully contribute to the mission and to grow personally.

Current open positions range from entry to senior levels, and are in the following worldwide locations: Raritan, NJ, Titusville, NJ, Spring House, PA, and La Jolla, CA in the USA; Beerse, Belgium, and Leiden, The Netherlands; and Shanghai, China.

The open positions can be found at http://careers.jnj.com/ by entering keyword “Statistics & Decision Sciences and Biostatistics” in the search box on the page.

Thank you for your consideration, and please feel free to forward this to any professionals in your network you feel may be a good match.

Guohua (James) Pan on behalf of the Statistics & Decision Sciences team at Johnson & Johnson

Tel: 908 927 7947; email: ipan3@its.jnj.com

Guohua (James) Pan on behalf of the Statistics & Decision Sciences team at Johnson & Johnson

Tel: 908 927 7947; email: ipan3@its.jnj.com

Guohua (James) Pan on behalf of the Statistics & Decision Sciences team at Johnson & Johnson

Tel: 908 927 7947; email: ipan3@its.jnj.com
Introduction to The Lotus Group

The Lotus Group is a professional recruiting firm which specializes in the pharmaceutical industry, specifically in statistician, statistical programming, data management and data science positions. We pride ourselves in providing excellent customer service and building solid relationships with our client companies so that we can provide the best opportunities for our candidates. As a result of our industry expertise and long-lasting relationships, we partner with the best pharmaceutical and biotech companies in the world and are able to recruit top quality talent for our clients. The Lotus Group team members are located in New Jersey, Boston, California, Pennsylvania and Chicago. We are effectively able to connect locally with candidates in strong pharmaceutical markets. We also regularly attend statistical and industry-related conferences, so we stay well-informed and well-connected.

We are hosting a career session (Session 31) at the 2018 Symposium in New Jersey on Friday 6/15 at 2:10-3:50pm ET. We welcome everyone to join us and chat with our distinguished and influential panel speakers:

Joshua Chen, Global Head of Biostatistics and Programming at Sanofi Pasteur
Lisa Chiacchierini, Associate VP of Statistics at Merck
Laura Meyerson, Retired VP of Biometrics from Biogen
Lei Wang, President and Managing Partner at The Lotus Group
Ouhong Wang, VP of Statistics at Vertex

The topic of our discussion is:

Effective Communication

We communicate every day of our lives, but do we think about the message(s) we are sending, do we choose our words carefully so as to clearly convey our message without ambiguity and to the right audience (without assuming they either don't know anything about what we are saying or they already have certain assumption which in both cases is often not true). Communicate is particularly important for statisticians, who often need to discuss and interpret very complex information to an audience who "only wants to know the p-value" or "whether there are more than 10% of Adverse Event X". Communication training probably exists at every company, and scientific communication training also exists at many, but this session will go even further not only in terms of how to clearly convey very complex interpretation of data to an audience, answer questions in a way that does not offend anyone (i.e., answer objectively using the data as a reference rather than using our subjective opinions) and, lastly, to be very clear and anticipate questions from an audience who may be particularly naive when it comes to data or, on the other hand, very savvy. The goal of this session is to prepare the presenter to come away with the skills that will work for any type of statistical presentation and, of course, make you even more attractive to upper management and/or future employees.
TechData Service Company

TechData is a clinical research and consulting service company. Our knowledge of and focus on clinical trial studies means that we are highly skilled at providing high quality and expertise service in the pharmaceutical industry. Pharmaceutical and biotech companies rely on our talented biostatistics teams and resource for their R&D projects and personnel requirements.

TechData has focused energies exclusively on clinical trials, building a reputation of excellence among leading R&D organizations. With an extensive network of talented biostatisticians and statistical programmers, TechData is capable of providing highest quality service and introducing the brightest professionals to companies seeking biostatistics services.

Since 2000, TechData has reflected its values and high standards of professionalism with customer aligned service. We work with our clients for a long-term "win-win" partnership relationship. We value our long-term relationship and understand their changing priorities in the context of individual client needs and industry trends. We have built many biostatistics teams consisting of hundreds of biostatisticians and programmers supporting nearly 100 sponsors.

For available positions at TechData, interested candidates please send resumes to Ju Zhang, Vice President at ju.zhang@techdataservice.com.

Director of Biostatistics

- Provide statistical support to Clinical Development Plan for multiple compounds
- Act as biostatistics Rep in pre-IND and NDA/BLA activities
- Provide statistical input to study protocols, develop statistical analysis plan

Biostatistician

- Work with clinical study teams on study design, development and review of study protocols.
- Develop and review SAP, including Table/listing/figure shells, and final study report.
- Work with the project statistician on design/review of case report form.

Statistical Programmer

- Providing statistical programming and validation support for clinical study reports
- Overseeing and coordinating programming activities
- Producing analysis datasets, statistical tables, figures, listings, Integrated Summaries of Safety (ISS), Integrated Summaries of Efficacy (ISE), electronic submissions and other internal and external requests (e.g., publications).
- Developing and maintaining SOPs, SWPs and other related technical documents, providing input to the Database and CRF Development, creating edit check programs.
AbbVie is a global, research-driven biopharmaceutical company committed to developing innovative advanced therapies for some of the world’s most complex and critical conditions. The company’s mission is to use its expertise, dedicated people and unique approach to innovation to markedly improve treatments across four primary therapeutic areas: immunology, oncology, virology and neuroscience. In more than 75 countries, AbbVie employees are working every day to advance health solutions for people around the world. For more information about AbbVie, please visit us at www.abbvie.com. Follow @abbvie on Twitter, Facebook or LinkedIn.
Statistics

The latest release of SAS/STAT® is now available. SAS/STAT 14.3 enriches numerous analyses and adds one more procedure to your toolkit.

**SAS/STAT 14.3 Highlights**
- Causal mediation analysis.
- Compartmental models for pharmacokinetic analysis.
- Fast quantile process regression.
- Cause-specific proportional hazards analysis for competing-risks data.
- Variance estimation by the bootstrap method for survey data analysis.

**Recent SAS/STAT Additions**
- Generalized additive models by penalized likelihood estimation.
- Two-stage fully efficient fractional imputation and fractional hot-deck methods for survey data.
- Estimation of causal treatment effects.
- Weighted GEE methods for longitudinal data analysis.
- Time-dependent ROC curves for Cox regression.

Learn more
support.sas.com/statnewreleases

© 2018 SAS Institute Inc. All rights reserved. GH32016 (0818)
THE SCIENCE of POSSIBILITY

Vertex creates new possibilities in medicine to cure diseases and improve people’s lives.

We work with leading researchers, doctors, public health experts and other collaborators who share our vision for transforming the lives of people with serious diseases, their families and society.

Vertex is hiring! For career opportunities, please visit https://www.vrtx.com/working-here

Vertex and the Vertex triangle logo are registered trademarks of Vertex Pharmaceuticals Incorporated. ©2018 Vertex Pharmaceuticals Incorporated | V008-US-00-0497/66/2018
Announcement: 2019 ICSA Applied Statistics Symposium

You are cordially invited to the 2019 ICSA Applied Statistics Symposium in Raleigh!

The ICSA Applied Statistics Symposium will be held from Sunday, June 9 to Wednesday, June 12, 2019 at the Raleigh Convention Center, North Carolina. Please send any inquiry to Dr. Wenbin Lu (wlu4@ncsu.edu).

Raleigh is the capital and second largest city in North Carolina, and is known as the "City of Oaks" for its many oak trees, which line the streets in the heart of the city. Raleigh is the home to North Carolina State University (NCSU) and is part of the Research Triangle Park (RTP) area, together with Durham (home of Duke University) and Chapel Hill (home of the University of North Carolina at Chapel Hill). Raleigh is also home to numerous cultural, educational, and historic sites, including the Duke Energy Center for the Performing Arts (home for North Carolina Symphony and Carolina Ballet), Museums of history and natural sciences, and Marbles Kids Museum. Details are forthcoming on the symposium website.

Call for Invited Session Proposals

We welcome your invited session proposals. The invited sessions will be processed through program committee. If you plan to organize an invited session, please communicate with one of the program committee members. An invited session will be 100 minutes with 4 speakers or 3 speakers and one discussant. A proposal includes 1) session title, 2) organizer, 3) session chair, 4) list of speakers and discussant. It is required to confirm all speakers' availability before submitting a proposal. There is a one-talk rule for speakers, but one can serve as a discussant in another invited session while speaking in an invited or contributed session. The deadline for the invited session proposal is November 16, 2018.

Call for Student Paper Award Applications

Up to eight student award winners (five Student Travel Awards, one Jiann-Ping Hsu Pharmaceutical and Regulatory Sciences Student Paper, and possible two ASA Biopharmaceutical Awards) will be selected. Each winner will receive a plaque or certificate, an award for travel and registration reimbursement up to $1,000 or a cash award of $550, whichever is bigger, as well as free registration for a short course. The deadline for applications is March 15, 2019.

Call for Short Course Proposals

Anyone who is interested in giving a one-day or half-day short course at 2019 ICSA Applied Statistics Symposium is welcome to submit a short-course proposal to Dr. Yonggang Yao (Yonggang.Yao@sas.com). The submission deadline is December 31, 2018.

Executive Committee

- Wenbin Lu, Chair of Organizing Committee, North Carolina State University
- Donglin Zeng, Program Committee Chair, University of North Carolina at Chapel Hill
- Kai Zhang, Program Book Chair, University of North Carolina at Chapel Hill
- Qing Yang, Local Committee Chair, Duke University
- Shu Yang, Treasurer, North Carolina State University
- Luo Xiao, Student Paper Award Chair, North Carolina State University
- Yonggang Yao, Short Course Chair, SAS Inc.
- Haoda Fu, Fund Raising Chair, Eli Lilly and Company
- Mengling Liu, Strategic Advisor, New York University
The 2019 ICSA China Conference
July 1–July 4, 2019, Tianjin, China

The 2019 ICSA China Conference will take place, July 1–July 4, 2019, in Tianjin, China. The Conference will be organized jointly by the International Chinese Statistical Association (ICSA), Nankai University and Shanghai Jiaotong University. It will be held on the campus of Nankai University.

The objective of the conference is to provide a platform for the exchange of recent research and developments in modern statistical methods, to create collaboration opportunities and to identify new directions for further research.

All sessions in the conference will be invited sessions. The Invited sessions will be organized through program committee members. We sincerely welcome ICSA members to propose invited sessions. Please contact one of the program committee members if you plan to organize an invited session. An invited session will be 90 minutes and include 4 speakers or 3 speakers and 1 discussant.

Program Committee:

Zhezhen Jin (Chair), Columbia University, New York, NY, USA
Mingyao Ai, Peking University, Beijing, China
Tianxi Cai, Harvard University, Boston, MA, USA
Ming-Hui Chen, University of Connecticut, Storrs, CT, USA
Luyan Dai, Harbour Biomed, Shanghai, China
Jianing Di, Janssen R&D, China
Peisong Han, University of Michigan, Ann Arbor, MI, USA
Mei-Ling Ting Lee, University of Maryland, Baltimore, MD, USA
Gang Li, Janssen Research and Development, Raritan, NJ, USA
Gang Li, University of California at Los Angeles, Los Angeles, CA, USA
Hongzhe Li, University of Pennsylvania, Pennsylvania, PA, USA
Dennis K.J. Lin, Pennsylvania State University, University Park, PA, USA
Yuanyuan Lin, Chinese University of Hong Kong, Hong Kong, China
Aiyi Liu, National Institute of Health, Washington D.C., USA
Mengling Liu, New York University, New York, NY, USA
Min-Qian Liu, Nankai University, Tianjin, China
Tao Liu, Brown University, Providence, RI, USA
Xuewen Lu, University of Calgary, Alberta, Canada
Ying Lu, Stanford University, Stanford, CA, USA
Changxing Ma, State University of New York at Buffalo, Buffalo, NY, USA
Robert Mee, University of Tennessee, Knoxville, TN, USA
Mounir Mesbah, Université Pierre et Marie Curie, P6, Paris, France
Taesung Park, Seoul National University, Seoul, Korea
Annie Qu, University of Illinois at Urbana–Champaign, Champaign, IL, USA
(Tony) Jianguo Sun, University of Missouri, Columbia, MO, USA
Yifei Sun, Columbia University, New York, NY, USA
Niansheng Tang, Yunnan University, Kunming, Yunnan Province, China
Haonan Wang, Colorado State University, Collins, CO, USA
Liqun Wang, University of Manitoba, Winnipeg, Manitoba, Canada
Pei Wang, Icahn School of Medicine at Mount Sinai, New York, NY, USA
Yazhen Wang, University of Wisconsin--Madison, Madison, WI, USA
Hulin Wu, University of Texas, Houston, TX, USA
Xiangrong Yin, University of Kentucky, Lexington, KY, USA
Ming Yuan, Columbia University, New York, NY, USA
Chunming Zhang, University of Wisconsin--Madison, Madison, WI, USA
Heping Zhang, Yale University, New Haven, CT, USA
Jin-Ting Zhang, Singapore National University, Singapore
Tracy Zhang, Biostatistics Group Head, Oncology China, Beijing, China
Linda Zhao, University of Pennsylvania, Philadelphia, PA, USA
Bojuan Zhao, Tianjin University of Finance and Economics, Tianjin, China
Zhigen Zhao, Temple University, Pennsylvania, PA, USA
Changliang Zou, Nankai University, Tianjin, China

Local Organizing Committee:

Zhaojun Wang (Chair), Nankai University, Tianjin, China
Chengming Bai, Nankai University, Tianjin, China
Dong Han, Shanghai Jiao Tong University, Shanghai, China
Min-Qian Liu, Nankai University, Tianjin, China
Weidong Liu, Shanghai Jiao Tong University, Shanghai, China
Zixiong Ren, Nankai University, Tianjin, China
Guijun Yang, Tianjin University of Finance and Economics, Tianjin, China
Yongdao Zhou, Nankai University, Tianjin, China
Changliang Zou, Nankai University, Tianjin, China
**Scientific Program (June 14th - June 17th)**

---

**June 14 2:00 PM - 3:00 PM**

**Keynote Session I (Keynote)**
Room: Regency DEF  
Organizer: ICSA 2018 organizing committee.  
Chair: Aiyi Liu, NIH & President, ICSA.

2:00 PM Keynote lecture I: Reflections on Statistical Practice, Leadership, and Training  
*Lisa LaVange.* University of North Carolina at Chapel Hill

**June 14 3:20 PM - 5:00 PM**

**Session 1: Network Data Analysis (Featured)**  
Room: Brunswick A (120)  
Organizer: Cheng Yong Tang, Temple University  
Chair: Cheng Yong Tang, Temple University.

- Model-assisted design of experiments on social networks  
  *Edoardo Airoldi.* Harvard University
- Why Aren’t Network Statistics Accompanied By Uncertainty Statements?  
  *Eric Kolaczyk.* Boston University
- Empirical risk minimization for relational data  
  *Peter Orbanz.* Columbia University
- Floor Discussion.

**Session 2: Future of Statisticians in the pharmaceutical world – A panel Discussion with Leaders (Featured)**  
Room: Brunswick B (120)  
Organizer:  
Jason Liao, Merck & Co., Inc.  
Gang Li, Johnson & Johnson  
Zhaoling Meng, Sanofi  
Hong Tian, Johnson & Johnson.
Facilitator: Surya Mohanty, Johnson & Johnson.

3:20 PM **Panelists:**  
Jesse Berlin, Sr. Vice President, Epidemiology, Johnson & Johnson  
Bruce Binkowitz, Vice President, Biostatistics & Programming, Shionogi, Inc  
Ivan Chan, Vice President, Pipeline Statistics and Programming, Data and Statistical Sciences, Abbvie  
Eric Genevois-Marlin, Vice President, Head of Global Biostatistics & Programming, Sanofi  
Jose Pinheiro, Head, Statistical Modeling & Methodology, Johnson & Johnson  
Bill Wang, Executive Director, BARDS, Merck & Co., Inc.

**Session 3: Some recent development in large-scale statistical learning with complex data (Invited)**  
Room: Garden State B (70)

**Session 4: SIBS Sponsored Session: Methods for Big Data Analysis in Genomics and Biomedicine (Invited)**  
Room: Conference I  
Organizer: Hongzhe Li, University of Pennsylvania.  
Chair: Yu Zhang, Pennsylvania State University.

3:20 PM Uncovering the hierarchical conformation of topologically associating domains from Hi-C data  
*Yu Zhang¹, Lin An, Tao Yang and Qunhua Li².*  
¹Pennsylvania State University  
²University of Pennsylvania

3:45 PM Estimating the overall contribution of human microbiome to the risk of developing cancers based on prospective studies  
*Jianxin Shi.* National Cancer Institute

4:10 PM A geometric comparison of the rejection regions of association tests  
*Ian Barnett¹, Zhonghua Liu² and Xihong Lin².*  
¹University of Pennsylvania  
²Harvard University

4:35 PM A semi-supervised approach for predicting cell type/tissue specific functional consequences of non-coding variation using massively parallel reporter assays  
*Iuliana Ionita-Laza.* Columbia University

5:00 PM Floor Discussion.

**Session 5: Basket Trials, Umbrella Trials, Platform Trials and Other Master Protocols in the Clinical Development of Cancer Targeted Therapies (Invited)**  
Room: Garden State C (70)  
Organizer: Mary Zhao, Boehringer Ingelheim Pharmaceuticals, Inc.  
Chair: Mary Zhao, Boehringer Ingelheim Pharmaceuticals, Inc.

3:20 PM Overview of Basket and Umbrella Trials: Features, Challenges, and Examples  
*Lindsay Renfro.* Mayo Clinic
Session 6: Recent Work on Clinical Trial Design and Analysis (Invited)
Room: Salon AB
Organizer: Ying Lu, Stanford University.
Chair: Ying Lu, Stanford University.
3:20 PM The current status and challenges of clinical efficiency assessment in Traditional Chinese Medicine
Buoyan Liu. Chinese Academy of Traditional Chinese Medicine Sciences
3:45 PM A Randomized Controlled Clinical Trial for Traditional Chinese Medicine Treatment of Primary Insomnia
Shiyun Yan. Institute of Clinical Basic Medicine, China Academy of Chinese Medical Science
4:10 PM Inferring a null or non-null effect for binary endpoints in randomized blinded clinical trials
Li-Hsin Chien1, Hua Jin2, Jitendra Ganju3 and Ying Lu1. 1Department of Biomedical Data Science and Center for Innovative Study Design, Stanford University 2South China Normal University and Stanford University 3Consultant, San Francisco, CA, USA
4:35 PM Floor Discussion.

Session 7: Statistical inferences in some semi-parametric models (Invited)
Room: Conference B (65)
Organizer: Yixin Fang, New Jersey Institute of Technology.
Chair: Yixin Fang, New Jersey Institute of Technology.
3:20 PM Longitudinal Mixed Effects Model with Fused Lasso Regularization
Judy Zhong. NYU Langone Health
3:45 PM Estimating Time-Varying Treatment Effect in Presence of Unmeasured Confounders
Haiqun Lin1 and Robert Rosenheck. 1Yale University
4:10 PM Online clustering using Gaussian mixture models
Lei Yang. New York University
4:35 PM Single-index model for inhomogeneous spatial point processes
Ji Meng Loh1 and Yixin Fang. 1New Jersey Institute of Technology
5:00 PM Floor Discussion.

Session 8: Recent Advances in Dose Response Clinical Trials (Invited)
Room: Conference A (65)
Organizer: Naitee Ting, Boehringer Ingelheim Pharmaceuticals, Inc.
Chair: Naitee Ting, Boehringer Ingelheim Pharmaceuticals, Inc.
3:20 PM A Cautionary Note When a Dose Ranging Study is Used for Proving the Concept
Qiqi Deng1, Kun Wang2, Xiaofei Bai1 and Naitee Ting3. 1Boehringer Ingelheim Pharmaceuticals, Inc. 2Shenzhen Middle School
3:45 PM Dynamic Development Paths for Expanding a Proof-of-Concept (PoC) Study to Explore Dose Range
Qiqi Deng, Xiaofei Bai and Naitee Ting. Boehringer Ingelheim Pharmaceuticals, Inc.
4:10 PM Phase II dose spacing based on PK recommendations
Junxian Geng, Shihai Lu and Naitee Ting. Boehringer Ingelheim Pharmaceuticals, Inc.
4:35 PM Discussant: Weidong Zhang, Pfizer Inc.
5:00 PM Floor Discussion.

Session 9: Statistics and machine learning at the frontier of data-driven application and technology (Invited)
Room: Garden State A (50)
Organizer: Donghui Yan, University of Massachusetts Dartmouth.
Chair: Donghui Yan, University of Massachusetts Dartmouth.
3:20 PM Reinforcement learning for taxi driver dispatching
Zhiwei Qin. DiDi Research America
3:45 PM Extracting data from tables and charts in natural document formats
David Rosenberg. Bloomberg / New York University
4:10 PM Data Science in e-commerce
Donghui Yan. University of Massachusetts Dartmouth
4:35 PM Discussant: Yuandong Tian, Facebook AI Research
5:00 PM Floor Discussion.

Session 10: Data-Driven Decision Making and Objective Statistical Modeling with Big Data Applications (Invited)
Room: Salon CD
Organizer: Yuhong Yang, University of Minnesota.
Chair: Wei Qian, University of Delaware.
Ching-Kang Ing. National Tsing Hua University
3:45 PM A New Information Criterion for Model Selection
Jie Ding1, Vahid Tarokh1 and Yuhong Yang2. 1Duke University 2University of Minnesota
4:10 PM Estimation and Optimization of Composite Outcomes
Daniel Luckett1, Eric Laber2 and Michael Kosorok3. 1University of North Carolina at Chapel Hill 2North Carolina State University
4:35 PM Efficient Online Bandit Multiclass Learning with sqrt(T) Regret
Chicheng Zhang. Microsoft Research
5:00 PM Floor Discussion.

Session 11: Dynamic Modeling and Machine Learning for Big Data from Large Cohort Studies (Invited)
Room: Conference C (65)
Organizer: Colin Wu, National Heart, Lung and Blood Institute.
Chair: Colin Wu, National Heart, Lung and Blood Institute.
3:20 PM Nonparametric Estimation of Risk Tracking Indices for Longitudinal Studies
♦ Xin Tian and Colin Wu. National Heart, Lung and Blood Institute
3:45 PM Cardiovascular Health Across the Lifespan
♦ Colin Wu¹ and Norrina Allen². ¹National Heart, Lung and Blood Institute ²Northwestern University
4:10 PM Local Box-Cox transformation on time-varying parametric models for smoothing estimation of conditional CDF with longitudinal data
♦ Mohammed Chowdhury³, Colin Wu² and Reza Modarres³. ³Kennesaw State University ²National Heart, Lung and Blood Institute ³George Washington University
4:35 PM Discussant: Zhaohai Li, George Washington University
5:00 PM Floor Discussion.

June 15 8:30 AM -9:45 AM

Opening Ceremony and Keynote Session II (Keynote)
Room: Regency DEF
Organizer: ICSA 2018 organizing committee.
Chair: Gang Li, Executive Director, ICSA.
8:35 AM Opening addressing
8:45 AM Keynote lecture II: The Changing Landscape of Drug Development: To Adapt or Not to Adapt
Dionne Price. Office of Biostatistics, Office of Translational Sciences Center for Drug Evaluation and Research (CDER), FDA

June 15 10:00 AM - 11:40 AM

Session 12: Deep learning and its mechanism (Featured)
Room: Brunswick A (120)
Organizer: Donghui Yan, University of Massachusetts Dartmouth
Aiyou Chen, Google, LLC.
Chair: Donghui Yan, University of Massachusetts Dartmouth
Aiyou Chen, Google, LLC.
- A Unified View of Deep Generative Models
  Eric Xing. Carnegie Mellon University
- A Geometric View to Optimal Transportation and Generative Model
  Xianfeng Gu. Stony Brook University
- Analysis on gradient descent dynamics in Deep Neural Network
  Yuandong Tian. Facebook AI Research
- Floor Discussion.

Session 13: Making sense of data for business: the role of statistics (Invited)
Room: Regency C (80)
Organizer: Dungang Liu, Lindner College of Business, University of Cincinnati.
Chair: Yichen Qin, Lindner College of Business, University of Cincinnati.
10:00 AM Multilevel Log-Gaussian Cox Process Modelling for Structured Temporal Point Processes with Applications in Stock Market Trading
Yongtao Guan. University of Miami
10:25 AM Learning heterogeneity in causal inference using sufficient dimension reduction
Wenbo Wu. University of Oregon
10:50 AM Threshold Factor Models for High-Dimensional Time Series
Xialu Liu. San Diego State University
11:15 AM Structural Break Detection in Financial Durations
Yaohua Zhang. Vertex Pharmaceuticals
11:40 AM Floor Discussion.

Session 14: Challenges and Innovations in Modeling Large-Scale Imaging Data (Invited)
Room: Garden State B (70)
Organizer: Fengqing (Zoe) Zhang, Drexel University.
Chair: Jiangtao Gou, Fox Chase Cancer Center.
10:00 AM A time-varying AR coefficient model of functional near-infrared spectroscopy data
♦ Marco Benedetti and Timothy Johnson. University of Michigan
10:25 AM Can neuroimaging diagnose neurodegenerative diseases?
♦ Lei Wang¹, Karteeck Popuri, Rakesh Balachandar, Kathryn Alpert, Donghuan Lu, Mahadev Bhalla, Ian Mackenzie, Robin Hsiung, Veronika Hanco, Kristen Warren, Konstantinos Arfanakis, Julie Schneider, David Bennett and Mirza Faisal Beg. ¹Feinberg School of Medicine, Northwestern University
11:15 AM Joint Modeling of Multimodal Neuroimaging Signatures of PTSD
♦ Fengqing Zhang and Xin Niu. Drexel University
11:40 AM Floor Discussion.

Session 15: Statistical advance in functional genomics (Invited)
Room: Garden State A (50)
Organizer: Qunhua Li, Department of Statistics, Penn State University.
Chair: Audrey Fu, Department of Statistics, University of Idaho.
10:00 AM Annotation Regression for Genome-Wide Association Studies

- Sungyoung Shin1 and Sunduz Keles2. 1University of Texas at Dallas 2University of Wisconsin-Madison

10:25 AM Learning causal biological networks with the principle of Mendelian randomization

- Audrey Fu and Md. bahadur Badsha. University of Idaho

10:50 AM Condition-adaptive fused graphical lasso: an adaptive procedure for inferring condition-specific gene co-expression network

- Qunhua Li3, Yafei Lyu1, Lingzhou Xue1, Feipeng Zhang1, Laura Saba and Katerina Kechris. 1Pennsylvania State University

11:15 AM Novel method for inferring genes that escape X chromosome inactivation from RNA seq data

- Dajiang Liu. Pennsylvania State University

11:40 AM Floor Discussion.

Session 16: Innovative Statistical Designs in Early Oncology Drug Development (Invited)
Room: Conference A (65)
Organizer: Nicole Li, Merck & Co., Inc.
Chair: Meihua Wang, Merck & Co., Inc.

10:00 AM Early oncology drug development: can we achieve more with less?

- Rong Liu. Bayer

10:25 AM Simulation Studies of Two Dose Escalation Methods for Oncology Drug Combination Therapies

- Jing Hu3, Ron Yu1 and Kc Huang2. 1Gilead Sciences 2Genentech

10:50 AM A time-to-event early oncology dose-finding framework for fast decision-making and incorporation of late-onset toxicity

- Zhen Zeng and Cong Chen. Merck & Co., Inc.

11:15 AM Benefits of Borrowing Information in Oncology Basket Trial

- Fang Liu and Cong Chen. Merck & Co., Inc.

11:40 AM Floor Discussion.

Session 17: Recent Advances in Nonparametric Methods (Invited)
Room: Regency A (80)
Organizer: Mei-Ling Ting Lee, University of Maryland, College Park.
Chair: Aiyi Liu, National Institutes of Health.

10:00 AM Rank-Based Tests for Clustered Data with R package clusrank

- Yajing Jiang1, Jun Yan2, Xin He3, Mei-Ling Ting Lee3 and Bernard Rosner4. 1University of Connecticut 2University of Maryland, College Park 4Harvard University

10:25 AM A log rank test for clustered data with informative within-cluster group size

- Somnath Datta1, Mary Gregg2 and Doug Lorenz2. 1University of Florida 2University of Louisville

10:50 AM Evaluating accuracy of diagnostic tests without conditional independence assumption

- Larry Tang. George Mason University

11:15 AM The noise barrier and the large signal bias of the Lasso and other convex estimators

- Pierre Bellec. Rutgers University

11:40 AM Floor Discussion.

Session 18: Advancing Complex Innovative Clinical Trial Design (Invited Panel)
Room: Brunswick B (120)
Organizer: Fanni Natanegara, Eli Lilly and Company.
Facilitator: Freda Cooner, Sanofi.

10:00 AM Panelists:

- Danise Subramaniam, Eli Lilly & Company
- Cristina Mayer, Johnson & Johnson
- Dionne Price, FDA

Session 19: Real-World Evidence: Moving from Data to Knowledge and Decision Making (Invited)
Room: Conference B (65)
Organizer: Xuanyao He, Eli Lilly and Company.
Chair: Xuanyao He, Eli Lilly and Company.

10:00 AM Transforming Real World Data into Scientific Evidence for the Regulatory Decision-Making

- Vandana Mukhi. Center for Devices and Radiological Health, FDA

10:25 AM Treatment decision using causal survival forests

- Yifan Cui1, Michael Kosorok1, Stefan Wager and Ruqing Zhu. 1University of North Carolina at Chapel Hill

10:50 AM Diagnosis of diabetic retinopathy using medical images and deep learning methods

- Xuanyao He and Haoda Fu. Eli Lilly and Company

11:15 AM Floor Discussion.

Session 20: The role of real world evidence in the development of medical treatment (Invited)
Room: Conference I (80)
Organizer: Xiwu Lin, Janssen Research and Development, LLC.
Chair: Jiayu Fu, Janssen China Research and Development.

10:00 AM Use of Real World Evidence for premarking evaluation of medical devices

- Nelson T. Lu. FDA

10:25 AM Applying propensity score adjustment in synthesizing real-world data in regulatory decision making

- Chenguang Wang. Johns Hopkins University

10:50 AM Weighting Strategies for Comparative Effectiveness Research using Observational Data


11:15 AM Discussant: Xiwu Lin, Janssen Research and Development, LLC

11:40 AM Floor Discussion.
Session 21: Statistical methods and applications in mobile health *(Invited)*  
Room: Regency B (80)  
Organizer: Chongzhi Di, Fred Hutchinson Cancer Research Center.  
Chair: Sisheng Liu, Fred Hutchinson Cancer Research Center.

10:00 AM Partially Observed Dynamic Models for Tracking Therapeutic and Engagement Outcomes  
♦ Zhenke Wu¹, Walter Dempsey² and Susan Murphy³.  
¹Department of Biostatistics, University of Michigan  
²Department of Statistics, Harvard University  
³Department of Statistics and of Computer Science, Harvard University  

10:30 AM Compositional and functional data analysis methods for accelerometry data  
♦ Chongzhi Di and Yifan Zhu. Fred Hutchinson Cancer Research Center

10:50 AM Clustering functional data with application to electronic medication adherence monitoring in HIV prevention trials  
♦ Yifan Zhu, Chongzhi Di and Ying Qing Chen. Fred Hutchinson Cancer Research Center

11:15 AM Discussant: Vadim Zipunnikov, Johns Hopkins University

11:40 AM Floor Discussion.

---

Session 22: Recent Developments in Risk Prediction Models *(Invited)*  
Room: Brunswick C (80)  
Organizer: Danping Liu, National Cancer Institute.  
Chair: Danping Liu, National Cancer Institute.

10:00 AM Dementia risk prediction in preclinical stage  
Zheyu Wang. Johns Hopkins University

10:25 AM Risk Prediction for Heterogeneous Populations with Application to Hospital Admission Prediction  
♦ Jared Huling¹, Menggang Yu, Muxuan Liang and Maureen Smith.  
¹Ohio State University

10:50 AM A Joint Model for Partial Peer Network Data with Informative Cluster Size  
♦ Danping Liu¹, Patrick Coyle², Denise Haynie³, Bruce Simons-Morton⁴ and Joe Bible¹.  
¹National Cancer Institute  
²Temple University  
³Eunice Kennedy Shriver National Institute of Child Health and Human Development  
⁴Clemson University

11:15 AM Floor Discussion.

---

Session 23: classification for high dimensional complex data *(Invited)*  
Room: Garden State C (70)  
Organizer: Ning Hao, University of Arizona.  
Chair: Han Xiao, Rutgers University.

10:00 AM Sparse quadratic classification rules via linear dimension reduction  
♦ Irina Gaynanova and Tianying Wang. Texas A & M University

10:25 AM Dynamic linear discriminant analysis in high dimensional space  
Binyan Jiang. The Hong Kong Polytechnic University

10:50 AM Quadratic discriminant analysis for high-dimensional data  
♦ Yingli Qin¹, Yilei Wu and Mu Zhu¹.  
¹University of Waterloo

11:15 AM Sparse Tensor Additive Learning for Click-Through-Rate Prediction  
Will Wei Sun. University of Miami

11:40 AM Floor Discussion.

---

Room: Conference C (65)  
Organizer: Sammy Yuan, Merck & Co., Inc..  
Chair: Christopher Assaid, Merck & Co., Inc..

10:00 AM Bayesian Approaches for individualized benefit-risk assessment  
♦ Yueqin Zhao¹, Ram Tiwari, Jyoti Zalkikar and Siqi Cui.  
¹U.S. Food and Drug Administration

10:25 AM Individual Benefit Risk Assessment: Overview and Proposal  
♦ Sammy Yuan and Shahruil Mt-Isa . Merck & Co., Inc.

10:50 AM Discussant: Bill Wang, Merck & Co., Inc., U.S.  
Discussant: Janet Wittes, Statistics Collaborative, Inc.

11:40 AM Floor Discussion.

---

June 15 1:00 PM -2:00 PM

Keynote Session III *(Keynote)*  
Room: Regency DEF  
Organizer: ICSA 2018 organizing committee.  
Chair: Minge Xie, Rutgers University.

1:00 PM Keynote lecture III: How good is your selected subgroup?  
Xuming He. University of Michigan

---

June 15 2:10 PM -3:50 PM

Session 25: New methods and challenges for analyzing of process data *(Featured)*  
Room: Brunswick A (120)  
Organizer: Yang Feng, Columbia University.  
Chair: Yi-Hsuan Lee, Educational Testing Service.

- New methods and challenges for analyzing process data in educational measurement  
Zhiliang Ying. Columbia University

- A joint Modeling Framework using responses and response times to track skill acquisition: Model Estimation and Application  
Shiyu Wang. University of Georgia

- Assessing 21st-century skills via human-human interactions: task design and psychometric challenges  
Jiangang Hao. Educational Testing Service

- Floor Discussion.
Session 26: *Biosimilar and bioequivalence (Invited)*
Room: Conference A (65)
Organizer: Yi Tsong, CDER, FDA.
Chair: Shemin-Chung Chow, CDER, FDA.

2:10 PM New Proposal for Equivalence Criteria in Bioequivalence Study with Binary Clinical Endpoint
Mengdie Yuan1, Jingyu Luan2 and Hui Sun. 1U.S. Food and Drug Administration

2:35 PM Sample Size Requirement for Analytical Biosimilarity Assessment
Tianhua Wang. U.S. Food and Drug Administration

3:00 PM Bioequivalence and Non-Inferiority Testing in The Presence of Subject-By-Treatment Interaction
Elena Rantou. U.S. Food and Drug Administration

3:25 PM Modified Wald Test for Reference Scaled Equivalence Assessment of Analytical Biosimilarity
Yu-Ting Weng1, Yi Tsong2, Meiuy Shen1 and Chao Wang1. 1U.S. Food and Drug Administration 2CDER, FDA

3:50 PM Floor Discussion.

Session 27: *Clinical trial design and statistical methods for cancer studies (Invited)*
Room: Conference B (65)
Chair: Fengqing Zhang, Drexel University.

2:10 PM A novel randomized Phase 2 trial design that constrains sample size by requiring a sufficient experimental response rate
Brian Egleston, Samuel Litwin, Stan Basickes and Eric Ross. Fox Chase Cancer Center

2:35 PM Sparse Minimum Discrepancy Approach to Sufficient Dimension Reduction with Simultaneous Variable Selection in Ultrahigh Dimension
Wei Qian1, Shanshan Ding1 and Dennis Cook2. 1University of Delaware 2University of Minnesota

3:00 PM Optimal timing of futility interim analyses
Dong Xi, Paul Gallo and David Ohslsen. Novartis

3:25 PM Hierarchical testing in a group sequential trial when the primary endpoint data become available earlier than the secondary endpoint data
Jiangtao Gou. Fox Chase Cancer Center

3:50 PM Floor Discussion.

Session 28: *Advance in Statistical Methods for Large and Complex Data (Invited)*
Room: Regency A (80)
Organizer: Dehan Kong, University of Toronto.
Chair: Dehan Kong, University of Toronto.

2:10 PM A Nonconvex approach to Sparse Generalized Eigenvalue Problem with Applications to Multivariate Statistics
Kean Ming Tan1, Zhaoran Wang, Han Liu and Tong Zhang. 1University of Minnesota

2:35 PM Online experimental design for automated high-throughput neural circuit mapping
Shizhe Chen. Columbia University

3:00 PM Concentration inequalities for time series
Fang Han. University of Washington

3:25 PM Statistical Inference for Online Learning and Stochastic Approximation via Hierarchical Incremental Gradient Descent
Weijie Su1 and Yuancheng Zhu. 1University of Pennsylvania

3:50 PM Floor Discussion.

Session 29: *Statistical Issues for Prognostic Biomarker Discovery and Risk Prediction in Cancer Genom (Invited)*
Room: Conference C (65)
Organizer: Li-Xuan Qin, Memorial Sloan Kettering Cancer Center.
Chair: Li-Xuan Qin, Memorial Sloan Kettering Cancer Center.

2:10 PM Measuring differential treatment benefit to evaluate predictive cancer biomarkers
Jaya Satagopan and Alexia Iasonos. Memorial Sloan Kettering Cancer Center

2:35 PM Gaussian process regression for survival time prediction with genome-wide gene expression
Aaron Molstad, Li Hsu and Wei Sun. Fred Hutchinson Cancer Research Center

3:00 PM Extreme Gradient Boosting for survival data
Xuefeng Wang1 and Zhenyu Zhang2. 1Moffitt Cancer Center 2Stony Brook University

3:25 PM Adjusting for Handling Effects in Microarray Data for Prognostic Biomarker Discovery and Survival Risk Prediction
Ai Ni1, Mengling Liu2 and Li-Xuan Qin1. 1Memorial Sloan Kettering Cancer Center 2New York University

3:50 PM Floor Discussion.

Session 30: *Revisiting challenges in time-to-event type of analysis in oncology trials (Invited)*
Room: Conference I (80)
Organizer: Li Li, Eli Lilly and Company.
Chair: Hong Wang, Eli Lilly and Company.

2:10 PM A Bayesian Joint Model of Recurrent Events and a Terminal Event
Ming Wang1, Zheng Li2 and Vern M. Chinchilli2. 1Pennsylvania State University College of Medicine 2Pennsylvania State University

2:35 PM Cure Model in Cancer Immunotherapy
Shu Wang1, Chengxing (Cindy) Lu2 and Rong Liu2. 1University of Pittsburgh 2Bayer

3:00 PM Missing data sensitivity analysis for time-to-event data in clinical trials
Xuezhou Mao, Chuaping Yu, Li Qi and Lan Yu. Sanofi

3:25 PM Validity of Real-World Control Arms for Single-Arm Clinical Trials: An Oncology Case Study
Zhanglin Cui, Yanping Wang and Douglas Faries. Eli Lilly and Company

3:50 PM Floor Discussion.
Session 31: Effective Communication (Invited Panel)  
Room: Brunswick B (120)  
Organizer: Lei Wang, The Lotus Group, LLC.  
Facilitator: Helena Fan, The Lotus Group, LLC.  

2:10 PM Panelists:  
Lei Wang, The Lotus Group, LLC  
Joshua Chen, Sanofi Pasteur  
Laura Meyerson, Merck & Co., Inc.  
Ouhong Wang, Vertex Pharmaceuticals  

Session 32: Statistical Process Control (Invited)  
Room: Brunswick C (80)  
Organizer: Dong Han, Shanghai Jiao Tong University.  
Chair: Yanhong Wu, California State University Stanislaus.  

2:10 PM On optimal control chart of change-point detection for dependent observation sequence  
Dong Han. Shanghai Jiao Tong University  

2:35 PM Run Length Properties of The EWMA Control Chart With Changing Limits  
Gaofeng Lv. Shanghai Jiao Tong University  

3:00 PM A Distribution-Free Control Chart for Monitoring High-Dimensional Processes with Individual Observation  
♦ Jinyu Fan and Lianjie Shu. University of Macau  

3:20 PM A Hybrid Hierarchical Bayesian Model for Spatio-Temporal Surveillance Data  
Jian Zou. Worcester Polytechnic Institute  

3:50 PM Floor Discussion.  

Session 33: Decision making in early clinical development of new drugs (Invited)  
Room: Regency B (80)  
Organizer: Wenqiong Xue, Boehringer Ingelheim Pharmaceuticals, Inc..  
Chair: Rui Wu, Boehringer Ingelheim Pharmaceuticals, Inc..  

2:10 PM PKQT Model Selection, Diagnostics and Prediction for One Oncology Drug  
Shu Yang. Novartis  

2:35 PM Dose-finding designs in the presence of disease progression  
♦ Bin Cheng¹, Lucie Biard and Shing Lee. ¹Columbia University  

3:00 PM An ROC based approach to interim Go/No-Go decision-making in clinical development  
Lanju Zhang. Abbvie, Inc.  

3:25 PM Using Bayesian proof-of-concept (PoC) criteria to design randomized Phase II oncology trials stratified by biomarker status  
Yong Zhang. Eisai  

3:50 PM Floor Discussion.  

Session 34: Statistical inference in data science (Invited)  
Room: Garden State A (50)  
Organizer: Yichuan Zhao, Georgia State University.  
Chair: Jason Liao, Merck & Co., Inc..  

2:10 PM Robust Jackknife Empirical Likelihood for Non-smooth U-structure Equations  
Yongli Sang. University of Louisiana at Lafayette  

2:35 PM Empirical likelihood for the bivariate survival function under univariate censoring  
♦ Yichuan Zhao and Haitao Huang. Georgia State University  

3:00 PM Smooth and non-smooth functionals of the NPMLE under interval censoring  
Zhigang Zhang. Memorial Sloan Kettering Cancer Center  

3:25 PM Autoregressive Model for matrix valued time series  
Han Xiao. Rutgers University  

3:50 PM Floor Discussion.  

Session 35: Advances in the pediatric drug development with extrapolation (Invited)  
Room: Regency C (80)  
Organizer: Zhaoling Meng, Sanofi.  
Chair: Zhaoling Meng, Sanofi.  

2:10 PM Utilization of Adult Data in Designing Pediatric Pharmacokinetic Studies: How Much is Historical Adult Data Worth?  
♦ Chyi-Hung Hsu and Steven Xu. Janssen Research and Development, LLC  

2:35 PM Modeling and simulation for planning and assessing the PK/PD consistency and comparison between pediatric and adult data  
♦ Mei Zhang and Zhaoling Meng. Sanofi  

3:00 PM Borrowing information from adult studies into pediatrics trials: Bayesian approach and a case study  
Jingjing Ye. FDA  

3:25 PM Discussant: Zhaoling Meng, Sanofi  
3:50 PM Floor Discussion.  

Session 36: Model Informed Clinical Drug Development (Invited)  
Room: Garden State B (70)  
Organizer: Susan Wang, Boehringer Ingelheim Pharmaceuticals, Inc..  
Chair: QiQi Deng, Boehringer Ingelheim Pharmaceuticals, Inc..  

2:10 PM Model Informed regulatory decisions  
Jingyu Yu. U.S. Food and Drug Administration  

2:35 PM Connecting the Dots to Inform Clinical Decision Making: PKPD Modeling, Meta-analysis, and Clinical Trial Design Optimization  
♦ Qi Tang, Zhaoling Meng, Lei Ma, Dimple Patel and Christine Veyrat-Follet. Sanofi  

3:00 PM Method on the determination of the threshold value for a continuous biomarker  
♦ Jianan Hui and Bushi Wang. Boehringer Ingelheim Pharmaceuticals, Inc.  

3:25 PM Discussant: Ying Lu, Stanford University  
3:50 PM Floor Discussion.  

2018 ICSA Applied Statistics Symposium, New Brunswick, NJ, June 14-17
Session 37: Novel Applications of Statistical Learning Methods in Complex Biomedical Data (Invited)
Room: Garden State C (70)
Organizer: Jiwei Zhao, State University of New York at Buffalo.
Chair: Jiwei Zhao, State University of New York at Buffalo.
2:10 PM Clustering Data on the Sphere: State of the Art and a Poisson Kernel-Based Algorithm
Mariamthi Markatou, University at Buffalo
2:35 PM Identification of Optimal Biomarker Combination for Treatment-Selection in the Presence of Missing Data
Yang Huang and Xiao-Hua Zhou. 1Fred Hutchinson Cancer Research Center 2University of Washington
3:00 PM Risk screening for Alzheimer’s Disease progression with Volume under the ROC Surface
Yu Cheng. University of Pittsburgh
3:25 PM Point and Interval Estimations for Individualized MCID
Jiwei Zhao. State University of New York at Buffalo
3:50 PM Floor Discussion.

June 15 4:10 PM - 5:50 PM

Session 38: Statistical Models for Understanding HIV/AIDS Epidemics (Invited)
Room: Garden State B (70)
Organizer: Le Bao, Pennsylvania State University.
Chair: Le Bao, Pennsylvania State University.
4:10 PM Trial Designs for Evaluating Combination HIV Prevention Approaches.
Sayan Dasgupta, Ying Qing Chen, Lili Peng, Yixin Wang, and Thomas Fleming. 1Fred Hutchinson Cancer Research Center 2University of Washington
4:35 PM Size Estimation of People At High Risk for HIV Infections
Le Bao. Pennsylvania State University
5:00 PM Comparison between HIV routine testing data and sentinel surveillance data
Ben Sheng, Jeffrey Eaton, Kimberly Marsh, Mary Mahy and Le Bao. 1Pennsylvania State University 2Imperial College London 3UNAIDS
5:25 PM Spatial-temporal distribution of newly detected HIV/AIDS cases among women aged ≥15 years old in China, 2010-2016
Fangfang Chen. National Center for AIDS/STD Control and Prevention, China CDC
5:50 PM Floor Discussion.

Session 39: Logical Inference on Efficacy in Subgroups and Their Mixtures: Targeted Therapies (Invited)
Room: Regency A (80)
Organizer: Hong Tian, Johnson & Johnson.
Chair: Ying Wan, Johnson & Johnson.
4:10 PM Logic-respecting Efficacy Measures
Jason Hsu. Ohio State University
4:35 PM The Subgroup Mixable Estimation Principle
Haian Xu and Jason Hsu. 1Johnson & Johnson 2Ohio State University
5:00 PM Logical Inference on Efficacy in Subgroups and Their Mixtures:
Hong Tian and Jason Hsu. 1Janssen Research and Development, LLC 2Ohio State University
5:25 PM Discussant: Yi Liu, Takeda
5:50 PM Floor Discussion.

Session 40: The Jiann-Ping Hsu Invited Session on Biostatistical and Regulatory Sciences (Invited)
Room: Brunswick A (120)
Organizer: Lili Yu, Georgia Southern University.
Chair: Jingxian Cai, PPD.
4:10 PM Applications of Mediation analysis to Estimands in Clinical Trials
Jingxian Cai and Andrew Hartley. PPD
4:35 PM K-subtree Pseudo-likelihood method in Phylogenetic tree model
Weifeng Wang. University of Georgia
5:00 PM Inference on the phylogenetic tree sets
Yan Du. University of Georgia
5:25 PM Floor Discussion.

Session 41: Statistical Inference for Distributed Data and Online Streaming Data (Invited)
Room: Regency B (80)
Organizer: Xi Chen, New York University.
Chair: Xi Chen, New York University.
4:10 PM Quantile Regression for big data with small memory
Xi Chen, Weidong Liu and Yichen Zhang. 1New York University
4:35 PM Communication-Efficient Distributed Statistical Inference
Yun Yang, Jason Lee and Michael Jordan. 1Florida State University 2University of Southern California 3University of California, Berkeley
5:00 PM Distributed estimation of principal eigenspaces
Ziwei Zhu, Jianqing Fan, Kaizheng Wang and Dong Wang. 1Princeton University
5:25 PM Statistical Inference for Model Parameters in Stochastic Gradient Descent
Yichen Zhang, Xi Chen, Jason Lee and Xin T. Tong. 1New York University 2University of Southern California 3National University of Singapore
5:50 PM Floor Discussion.

Session 42: New developments in microbiome sequencing data modeling and analysis (Invited)
Room: Garden State C (70)
Organizer: Zhigang Li, Geisel School of Medicine at Dartmouth College.
Chair: Zhiguang Huo, University of Florida.
Session 43: Flexible modeling and inference for complex data
(Invited)
Room: Regency C (80)
Organizer: Yichao Wu, University of Illinois at Chicago.
Chair: Guan Yu, University at Buffalo.

4:10 PM A two-stage microbial association mapping framework with advanced FDR control
♦ Jiyuan Hu1, Hyunwook Koh2, Linchen He2, Huilin Li2, Menghan Liu2 and Martin Blaser2. 1NYU Langone Medical Center 2NYU School of Medicine

4:35 PM Optimal Error-in-Variable Compositional Regression for Next-Generation Sequencing Data
♦ Pixu Shi, Anru Zhang and Yichen Zhou. University of Wisconsin-Madison

5:00 PM A Random Effects Model for Large-Scale Multinomial Data
♦ Wei Lin and Jingru Zhang. Peking University

5:25 PM GLM-based latent variable ordination method
♦ Michael Sohn and Hongze Li. University of Pennsylvania

5:50 PM Floor Discussion.

Session 44: Statistical solutions to practical issues in educational testing (Invited)
Room: Garden State A (50)
Chair: Yi-Hsuan Lee, Educational Testing Service.

4:10 PM Best Linear Prediction of Test True Scores
Lili Yao. Educational Testing Service

4:35 PM Modeling Writing Processes Using Keystroke Logs
♦ Hongwen Guo, Mo Zhang, Paul Deane and Randy Bennett. Educational Testing Service

5:00 PM Latent Regression Model for Group Score Estimation: Model Selection and Model Reduction
Yue Jia. Educational Testing Service

5:25 PM Statistical Procedures for Detecting Large-Scale Unusual Similarity in Responses
Yi-Hsuan Lee. Educational Testing Service

5:50 PM Floor Discussion.

Session 45: New Topics in Statistical Learning (Invited)
Room: Conference A (65)
Organizer: Ganggang Xu, Binghamton University.
Chair: Wei Sun, University of Miami.

4:10 PM Individualized Treatment Rules with Reject and Refine Options
♦ Xingye Qiao1 and Haomiao Meng. 1Binghamton University

4:35 PM High-dimensional cost-constrained regression via non-convex optimization
♦ Guan Yu1, Haoda Fu2 and Yufeng Liu. 1State University of New York at Buffalo 2Eli Lilly and Company

5:00 PM Likelihood Ratio Test for Stochastic Block Models with Bounded Degrees
Zuofeng Shang. Indiana University-Purdue University Indianapolis

5:25 PM A Big Data Linear Regression Via A-optimal Subsampling
♦ Sean Peng and Fei Tan. Indiana University-Purdue University Indianapolis

5:50 PM Floor Discussion.

Session 46: Futility analysis and interim decision making (Invited)
Room: Conference B (65)
Organizer: Qiqi Deng, Boehringer Ingelheim Pharmaceuticals, Inc.
Chair: Miguel Garcia Jr, Boehringer Ingelheim Pharmaceuticals, Inc.

4:10 PM Extending a futility rationale to multiple endpoint settings
♦ Paul Gallo and Dong Xi. Novartis

4:35 PM Interim Futility Analysis for Longitudinal Data With Adaptive Timing and Error Rate Preservation
G. Cai. GlaxoSmithKline

5:00 PM Statistical Consideration when Combining Two Independent Trials in Interim Futility Analysis
♦ Dooti Roy1, Qiqi Deng2, Ming-Hui Chen2 and Ying Ying Zhang. 1Boehringer Ingelheim Pharmaceuticals, Inc. 2University of Connecticut

5:25 PM Futility and the DSMB: An Unholy Alliance?
Janet Wittes. Statistics Collaborative, Inc.

5:50 PM Floor Discussion.

Session 47: Recent Biomarker Developments with Applications in Brain Research and Disorders (Invited)
Room: Brunswick B (120)
Organizer: Rosanne Lane, Janssen Research and Development, LLC.
Chair: Rosanne Lane, Janssen Research and Development, LLC.

4:10 PM Constructing Concurrent Network of Biomarker Processes Using Dynamical Systems
♦ Ming Sun and Yuanjia Wang. University of Pennsylvania

4:35 PM Adjusting for Covariate Measurement Error in Failure Time Analysis under Competing Risks
♦ Carrie Caswell and Sharon Xie. University of Pennsylvania
5:00 PM Determine Appropriate Sample Size for a Biomarker Signature Discovery Problem Using Penalized Regression
  †Xiang Li, Hong Tian and Liang Xiu. Janssen Research and Development, LLC

5:25 PM Discussant: Yevgen Tymofyeyev, Janssen Research and Development, LLC

5:50 PM Floor Discussion.

Session 48: Collaboration Space between Biostatistics and Pharmacometrics (Invited)
Room: Conference C (65)
Organizer: Jose Pinheiro, Janssen.
Chair: Jose Pinheiro, Janssen Research and Development, LLC.

4:10 PM This isn’t peewee soccer - building teamwork between statisticians and pharmacometricians
  Bret Musser. Regeneron Pharmaceuticals, Inc.

4:35 PM Synergy between Pharmacometrics and Statistics in Drug Development Decision Making
  Liping Zhang. Janssen Research and Development, LLC

5:00 PM Discussant: David Ohlssen, Novartis.
Discussant: Robert Bies, SUNY Buffalo.

5:50 PM Floor Discussion.

Session 49: Study Designs for Phase I Oncology Dose-Finding Trials (Invited)
Room: Conference I (80)
Organizer: Mary Zhao, Boehringer Ingelheim Pharmaceuticals, Inc.
Chair: Zhichao Sun, Boehringer Ingelheim Pharmaceuticals, Inc.

4:50 PM Embracing model-based designs for dose finding trials
  Birgit Gaschler-Markefski. Boehringer Ingelheim Pharmaceuticals, Inc.

4:10 PM AAA: Triple-adaptive Bayesian designs for the identification of optimal dose combinations in dual-agent dose-finding trials
  †Yuan Ji1 and Jiaying Lyu2. 1University of Chicago 2Fudan University

4:30 PM Flexible early-phase design for combination therapies
  Nolan Wages. University of Virginia

5:10 PM Time-to-event Bayesian Optimal Interval Design to Accelerate Phase I Trials
  Ruitao Lin. University of Texas M.D. Anderson Cancer Center

5:30 PM Accuracy, Safety and Reliability of Novel Phase I Trial Designs
  †Ying Yuan2, Lei Nie2 and Heng Zhou1. 1University of Texas MD Anderson Cancer Center 2U.S. Food and Drug Administration

5:50 PM Floor Discussion.

Session 50: Design and execution of late stage oncology trials in modern cancer drug development (Invited)
Room: Brunswick C (80)
Organizer: Rong Liu, Bayer.
Chair: Rong Liu, Bayer.

4:10 PM A pitfall in monitoring survival trials under non-proportional hazards
  Xiaodong Luo. Sanofi

4:35 PM A Flexible Survival Model for Fitting Time to Event Data in Clinical Trials
  †Jason Liao1 and Frank Liu. 1Merck & Co., Inc.

5:00 PM Multiple Hypothesis Testing and Timing of Analyses in Immuno-oncology Trials
  †Christine Gause and Keaven Anderson. Merck Research Laboratories

5:25 PM Apply Bayesian Response Adaptive Design to Select Winner from Novel Cancer Treatment
  Xiaoling Wu. Celgene

5:50 PM Floor Discussion.

June 16 9:50 AM - 11:30 AM

Keynote Session IV (Keynote)
Room: Regency DEF
Organizer: ICSA 2018 organizing committee.
Chair: Heping Zhang, Yale University & President-elect, ICSA.

8:30 AM Keynote lecture IV: Statistical Inference With High-Dimensional Data
  Cun-Hui Zhang. Rutgers University

June 16 9:50 AM - 11:30 AM

Session 51: Challenges and advances in imaging-genetics (Featured)
Room: Garden State B (70)
Organizer: Gen Li, Department of Biostatistics, Columbia University.
Chair: Gen Li, Department of Biostatistics, Columbia University.

- Challenges and Progress in Imaging Genetic - Big Data Squared - Studies
  Heping Zhang. Yale University

- Bayesian nonparametric method for genetic dissection of brain activation region
  †Jian Kang1, Zhuxuan Jin2 and Tianwei Yu2. 1University of Michigan 2Emory University

- Relationships Between Brain Structural Connectome and Traits
  Zhengwu Zhang. University of Rochester

- Floor Discussion.

Session 52: Integrative analysis of multi-view data with applications to precision medicine (Invited)
Room: Garden State B (70)
Organizer: Gen Li, Department of Biostatistics, Columbia University.
Chair: Gen Li, Department of Biostatistics, Columbia University.

9:50 AM Joint Modeling of Multi-System Wearable Data
  Vadim Zipunnikov. Johns Hopkins Bloomberg School of Public Health
10:15 AM Integrating large-scale sequencing data for cancer classification  
Ronglai Shen. Memorial Sloan Kettering Cancer Center

10:40 AM Integrative clustering of multidimensional omics data  
*Mengyun Wu*, Sebastian Teran Hidalgo, Yang Li and Shuangge Ma. 1Yale University 2Renmin University of China

11:05 AM An Integrative Graphical Modeling Approach for Multiple Heterogeneous Omics Data  
*Annaliza Megillivray* and *George Michailidis*. 1University of Florida Informatics Institute 2University of Florida

11:30 AM Floor Discussion.

**Session 53: Real world evidence in the US health care system** (Invited)  
Room: Brunswick B (120)  
Organizer: Victoria Gamerman, Boehringer Ingelheim Pharmaceuticals, Inc.  
Chair: Nan Shao, Boehringer Ingelheim Pharmaceuticals, Inc.

9:50 AM Successfully Integrate RWE in Clinical Processes in the Era of Big Data  
Kelly Zou. Pfizer Inc.

10:15 AM Statistical challenges and method-exploration for intermittent missing data occurring concurrently with right censoring: Quantitative illustration from a real-world study evaluating treatment patterns and unmet needs in adult patients with moderate-to-severe Atopic Dermatitis (AD)  

10:40 AM Illuminating variation in implantable cardiac device use and outcomes with billing claims data  
*Laura Hatfield*, Sharon-Lise Normand, Daniel Kramer and Rita Volya. 1Harvard Medical School 2Beth Israel Deaconess Medical Center

11:05 AM Real world challenges in the design of pragmatic trials  
Valentina Zabek. Boehringer Ingelheim Pharmaceuticals, Inc.

11:30 AM Floor Discussion.

**Session 54: Statistical Inference for Discrete and Categorial Data** (Invited)  
Room: Regency B (80)  
Organizer: Robin Gong, Harvard University.

Chair: Paul Edlefsen, Fred Hutchinson Cancer Research Center.

9:50 AM Evaluating vaccine efficacy variation by type of infecting pathogen using Dempster-Shafer (DS) Multinomial models.  
*Paul Edlefsen*, Robin Gong and Jan Hannig. 1Fred Hutchinson Cancer Research Center 2Harvard University 3University of North Carolina at Chapel Hill

10:15 AM Change Point Detection for Poisson Time Series Images with Applications to Astronomy and Astrophysics  
Thomas Lee. University of California, Davis

10:40 AM Poisson is enticing, but this preprocessed long-tail fish has fewer bones...  
Xiao-Li Meng. Harvard University

11:05 AM Floor Discussion.

**Session 55: Modern Statistical Development for Biomedical Big Data** (Invited)  
Room: Regency C (80)  
Organizer: Qingxia Chen, Vanderbilt University Medical Center.

Chair: Zhiqiang Huo, University of Florida.

9:50 AM Bayesian Large Scale Inference for Time to First Event of Multivariate Ordinal Survival Outcomes with Application to TennCare Cohort Study  
*Qingxia Chen*, David Schlueter, Christopher Fonnesbeck and Pingsheng Wu. 1Vanderbilt University Medical Center 2Vanderbilt University

10:15 AM Online Updating of Survival Analysis in the Big Data Setting  
*Jing Wu*, Elizabeth Schifano, Jan Yan, Yishu Xue and Ming-Hui Chen. 1University of Rhode Island 2University of Connecticut

10:40 AM Estimating Causal Log-Odds Ratio Using Case-Control Sample and Its Application in the Pharmaco-Epidemiology Study  
Donglin Zeng. University of North Carolina at Chapel Hill

11:05 AM Quantile Decision Trees and Forest with its application for predicting the risk (Post-Traumatic Stress Disorder) PTSD after experienced an acute coronary syndrome  
*Ying Wei*, Huichen Zhu and Ian Kronish. 1Columbia University

11:30 AM Floor Discussion.

**Session 56: New advanced methods in functional data analysis** (Invited)  
Room: Regency A (80)  
Organizer: Shujuie Ma, University of California, Riverside.

Chair: Li Cai, Soochow University.

9:50 AM Optimal weighting schemes for longitudinal and functional data  
*Xiaoke Zhang* and *Jane-Ling Wang*. 1George Washington University 2University of California, Davis

10:15 AM A simultaneous confidence band for variance function based on dense functional data  
Suojin Wang. Texas A & M University

10:40 AM Supervised Principal Component Regression for Functional Data with High Dimensional Predictors  
*Dehan Kong*, Xinyi Zhang and Qiang Sun. 1University of Toronto

11:05 AM Oracleally Efficient Estimation for Dense Functional Data With Holiday Effects  
Li Cai. Soochow University

11:30 AM Floor Discussion.

**Session 57: Dimension reduction, variable selection and their applications** (Invited)  
Room: Garden State A (50)  
Organizer: Xiangrong Yin, University of Kentucky.

Chair: Wenbo Wu, University of Oregon.
9:50 AM Sufficient Variable Selection Using Independence Measures for Continuous Responses
*Xiangrong Yin*\(^1\), Baoying Yang and Nan Zhang. 
\(^1\)University of Kentucky

10:15 AM Some Dimension Reduction Strategies for the Analysis of Survey Data
*Derek Young and Jiaying Weng*. University of Kentucky

10:40 AM Spatial envelope models
*Qin Wang*. Virginia Commonwealth University

11:05 AM Efficient integration of sufficient dimension reduction and prediction in discriminant analysis
*Xin Zhang*. Florida State University

11:30 AM Floor Discussion.

**Session 58: Analysis and computation for complex models**
*(Invited)*

Room: Brunswick C (80)
Organizer: Wentao Li, Newcastle University, UK.
Chair: Wentao Li, Newcastle University, UK.

9:50 AM Bayesian functional quantile regression
*Yusha Liu*\(^1\), Jeffrey Morris\(^2\) and Meng Li\(^1\). 
\(^1\)Rice University \(^2\)University of Texas M.D. Anderson Cancer Center

10:15 AM Validation of Approximate Bayesian Computation on Posterior Convergence
*Wentao Li*\(^1\) and Paul Fearnhead\(^2\). 
\(^1\)Newcastle University \(^2\)Lancaster University

10:40 AM Learning Temporal Evolution of Spatial Dependence
*Shiwei Lan*. Department of Computing + Mathematical Sciences, California Institute of Technology

11:05 AM Floor Discussion.

**Session 59: Hierarchical generalized linear models in practice** *(Invited)*

Room: Garden State C (70)
Organizer: Youngjo Lee, Seoul National University.
Chair: Ying Lu, Stanford University.

9:50 AM Optimality of the maximum h-likelihood estimators
*Youngjo Lee*. Seoul National University

10:15 AM H-likelihood Inference for Survival Models with Random Effects
*Il do Ha*. Pukyong National University

10:40 AM Power and false discovery rates of HGLMs and other statistical models analyzing RNAseq data
*Dongseok Choi*\(^1\), Maengseok Noh\(^2\), Jiwoong Kim\(^3\) and Youngjo Lee\(^4\). 
\(^1\)Oregon Health & Science University \(^2\)Pukyong National University \(^3\)Seoul National University

11:05 AM Data analysis using MDHGLM package in R
*Maengseok Noh*. Pukyong National University

11:30 AM Floor Discussion.

**Session 60: Building Successful Sponsor/CRO Relationship - A Biostatistical Perspective** *(Invited)*

Room: Conference A (65)
Organizer: Xiaohua Sheng, FMD K&L, Inc.
Chair: Xiaohua Sheng, FMD K&L, Inc.

9:50 AM Statistical Considerations: Best practices for Sponsor/CRO Relationships
*Patrick Larson*. Merck & Co., Inc.

10:15 AM Key Components for a Successful Sponsor/CRO Strategic Partnership
*Jianrong Li*. PPD

10:40 AM Functional Service Outsourcing Trends and Experience
*Tiepu Liu*. FMD K&L, Inc.

11:05 AM Discussant: Mei Wu, Sanofi

11:30 AM Floor Discussion.

**Session 61: Statistical and Regulatory Issues in Human Drug Abuse Studies** *(Invited)*

Room: Conference B (65)
Organizer: Qianyu Dang, U.S. Food and Drug Administration.
Chair: Steve Bai, U.S. Food and Drug Administration.

9:50 AM Statistical Issues in Human Drug Abuse Studies
*Qianyu Dang*. U.S. Food and Drug Administration

10:15 AM Statistical and Regulatory Issues in Human Drug Abuse Studies
*Catherine Mills and Denise Milovan*. Syneos Health

10:40 AM Overview of Human Abuse Potential Studies as Applied to Abuse Deterrent Assessment
*James Tolliver*. U.S. Food and Drug Administration

11:05 AM Floor Discussion.

**Session 62: Statistical Innovations in Big Data Analysis** *(Invited)*

Room: Conference C (65)
Chair: Yuping Zhang, University of Connecticut.

9:50 AM Parallel Markov Chain Monte Carlo Methods for Bayesian Analysis of Big Data
*Erin Conlon*. University of Massachusetts Amherst

10:15 AM More Efficient Estimation for Logistic Regression with Optimal Subsample
*Haiying Wang*. University of Connecticut

10:40 AM Joint Principal Trend Analysis for Longitudinal High-dimensional Data Integration
*Yuping Zhang*\(^1\) and Zhengqing Ouyang\(^2\). 
\(^1\)University of Connecticut \(^2\)The Jackson Laboratory for Genomic Medicine

11:05 AM On adaptive weighting of omics meta-analysis
*Yongseok Park*. University of Pittsburgh

11:30 AM Floor Discussion.
Conditional modeling of survival data with semi-competing risks

- Tonghui Yu and Liming Xiang. Nanyang Technological University

Approximate confidence distribution computing: An effective likelihood-free method with statistical guarantees

- Suzanne Thornton¹, Wentao Li and Minge Xie¹. Rutgers University ²Lancaster University

A Fast Score Test for Generalized Mixture Models

- Rui Duan¹, Yang Ning², Shuang Wang³, Lindsay Bruce⁴, Raymond Carroll⁴ and Yong Chen¹. ¹University of Pennsylvania ²Cornell University ³Columbia University ⁴Pennsylvania State University ²Texas A & M University

Generalized Integrative Principal Component Analysis for Multi-Type Data with Block-Wise Missing Structure

- Huichen Zhu¹, Gen Li¹ and Eric Lock². ¹Columbia University ²University of Minnesota

Floor Discussion.

June 16 1:00 PM -2:40 PM

Session 64: An introduction to BFF: new statistical inference tools for data science (Featured)

Room: Brunswick A (120)
Organizer: Minge Xie, Rutgers University.
Chair: Minge Xie, Rutgers University.

1:00 PM Could descendants of Jeffreys, Fisher and Neyman become best friends forever?

- Jan Hannig. University of North Carolina at Chapel Hill
Discussant: Xiao-Li Meng, Harvard University
Discussant: Thomas Lee, University of California, Davis
Discussant: Robin Gong, Harvard University
Floor Discussion.

Session 65: New Advances in Functional Data Analysis (Invited)

Room: Regency A (80)
Organizer: Pang Du, Virginia Polytechnic Institute and State University.
Chair: Pang Du, Virginia Polytechnic Institute and State University.

1:00 PM Principal Component Analysis for Functional Data on Riemannian Manifolds and Spheres

- Xiongtao Dai⁻¹ and Hans-Georg Mueller. ¹University of California, Davis

1:25 PM Covariance Estimation and Principal Component Analysis for Spatially Dependent Functional Data

- Yehua Li. University of California, Riverside

1:50 PM Bayesian Spline Smoothing with Ambiguous Penalties

- Xinlian Zhang, Gauri Datta, Ping Ma and Wenxuan Zhong. University of Georgia

2:15 PM Partially functional linear regression in high dimensions.

- Hao Zhang¹, Dehan Kong², Fang Yao and Kaijie Xue. ¹University of Arizona ²University of Toronto

2:40 PM Floor Discussion.

Session 66: SII Sponsored Invited Session: Modern Statistical Interface with Applications (Invited)

Room: Garden State B (70)
Organizer: Ming-Hui Chen, University of Connecticut.
Chair: Heping Zhang, Yale University.

1:00 PM Smoothing spline ANOVA for super-large samples: Scalable computation via rounding parameters

- Nathaniel Helwig¹ and Ping Ma². ¹University of Minnesota ²University of Georgia

1:25 PM Discovering Stock Chart Patterns by Statistical Estimation and Inference

- Hoang Tran¹ and Yiyan She². ¹Rho, Inc. ²Florida State University

1:50 PM MCMC algorithms for empirical Bayes analysis of rank data

- Vivekananda Roy¹, Arnab Laha and Somak Dutta. ¹Iowa State University

2:15 PM Dimension Reduction for Big Data

- Tonglin Zhang. Purdue University

2:40 PM Floor Discussion.

Session 67: Machine learning and causal inference with some industrial practices (Invited)

Room: Conference A (65)
Organizer: Aiyoun Chen, Google, LLC.
Chair: Aiyoun Chen, Google, LLC.

1:00 PM Automated Machine Learning

- Jin Cao. Bell Labs, Nokia

1:25 PM Application of Causal Bayesian Networks to Environmental Data

- Jie Hu, Carlos Carrion, Andrew Mcgowan and Megan Chen. The Climate Corporation

1:50 PM Hierarchical Modeling and Shrinkage for User Session Length Prediction in Media Streaming

- Zhen Zhu¹, Antoine Dedieu², Rahul Mazumder² and Hossein Vahabi³. ¹Pandora Media, Inc. ²Massachusetts Institute of Technology

2:15 PM Bias Correction For Paid Search In Media Mix Modeling

- Aiyoun Chen¹, David Chan, Michael Perry, Yuxue Jin, Haiting Sun¹, Nancy Wang and Jim Koehler. ¹Google, LLC

2:40 PM Floor Discussion.

Session 68: Learning from Big Data in Life Science and Translational Medicine (Invited)

Room: Regency C (80)
Organizer: Ying Lu, Stanford University.
Chair: Long Feng, Yale University.

1:00 PM Big-Data Analysis Points Toward A New Cancer Drug Discovery Method

- Bin Chen. Michigan State University

1:25 PM Mobile Health Apps Transforming Medical Research and Clinical Care: Using Apple’s new Research Kit for Asthma

- Pei Wang. Icahn School of Medicine at Mount Sinai

2018 ICSA Applied Statistics Symposium, New Brunswick, NJ, June 14-17
1:00 PM High-dimensional statistical inferences with over-identification: confidence set estimation and specification test
♦ Cheng Yong Tang\(^1\), Jinyuan Chang and Tong Tong Wu. \(^1\)Temple University

1:25 PM Functional Debias Estimation of Genetic Relatedness in High-dimensional Linear Models
♦ Wanjie Wang\(^1\), Zijian Guo\(^2\), Tony Cai\(^3\) and Hongze Li\(^2\). \(^1\)National University of Singapore \(^2\)Rutgers University \(^3\)University of Pennsylvania

1:50 PM Collaborative Spectral Clustering in Attributed Networks
Pengsheng Ji. University of Georgia

2:15 PM Graph-Based Two-Sample Tests for Discrete Data
♦ Jingru Zhang\(^1\) and Hao Chen\(^2\). \(^1\)Peking University \(^2\)University of California, Davis

2:40 PM Floor Discussion.

Session 72: Recent developments for causal effect estimation in observational studies (Invited)
Room: Garden State C (70)
Organizer: Kean Ming Tan, University of Minnesota.
Chair: Zhichao Jiang, Princeton University.

1:00 PM Kernel-based covariate functional balancing for observational studies
♦ Kwun Chuen Gary Chan\(^1\) and Raymond Wong. \(^1\)University of Washington

1:25 PM Robust Estimation of Propensity Score Weights via Subclassification
♦ Linbo Wang\(^1\), Xiao-Hua Zhou\(^2\) and Thomas Richardson\(^3\). \(^1\)Harvard University \(^2\)Peking University \(^3\)University of Washington

1:50 PM Multiply Robust Causal Inference With Double Negative Control Adjustment for Unmeasured Confounding
♦ Xu Shi\(^1\), Wang Miao\(^2\) and Eric Tchetgen Tchetgen\(^2\). \(^1\)Department of Biostatistics, Harvard University \(^2\)Wharton School, University of Pennsylvania

2:15 PM Adaptive Estimation in Structured Factor Models with Applications to Overlapping Clustering
Yang Ning. Cornell University

2:40 PM Floor Discussion.

Session 73: Challenges and Current Methods in Consumer Healthcare Studies (Invited)
Room: Brunswick C (80)
Organizer: Chunming Li, Pfizer Inc..
Chair: Chunming Li, Pfizer Inc..

1:00 PM Differences in the Application of Dental Pain Model across Sites and their Impact
♦ Rina Leyva and Chunming Li. Pfizer Inc.

1:25 PM Sample Size Evaluation in Bioequivalence Studies with a Replicated Crossover Design
♦ Jing Li and Chunming Li. Pfizer Inc.

1:50 PM Risk Modeling for Rx to OTC Switch to inform the selection of the threshold for rates misuse
Ching-Ray Yu. Pfizer Inc.
2:15 PM Statistical issues and considerations in equivalence assessment of SPF: in vivo and in vitro
  - Junhong Zhu\textsuperscript{1} and Eduardo Ravolo\textsuperscript{2}. 1Ipsen Biopharmaceuticals, Inc. 2Bayer

2:40 PM Floor Discussion.

Session 74: The Future of Big Data and Artificial Intelligence
(Invited Panel)
Room: Brunswick B (120)
Organizer: Kelly Zou, Pfizer Inc.
Facilitator: Kelly Zou, Pfizer Inc.
1:00 PM Panelists:
  - Aiyi Liu, National Institutes of Health
  - William Marder, IBM Watson Health
  - James Rosenberger, National Institute of Statistical Sciences
  - Cheng Zhang, Clover Health
  - Tian Zheng, Columbia University
  - Kelly Zou, Pfizer Inc

Session 75: Statistical Analysis of Textual Data (Invited)
Room: Conference B (65)
Organizer: Yunting Sun, Google, LLC.
Chair: Yunting Sun, Google, LLC.
1:00 PM Which Encoding is the Best for Text Classification in Chinese, English, Japanese and Korean?
  - Xiang Zhang. New York University
1:25 PM Rare Feature Selection in High Dimensions
  - Xiaohan Yan\textsuperscript{1} and Jacob Bien\textsuperscript{2}. 1Cornell University 2University of Southern California
1:50 PM Organic ranking of competing domains
  - Yunting Sun and Aiyou Chen. Google, LLC
2:15 PM Floor Discussion.

Session 76: Student Awards, Session II (ICSA Student Paper Awards)
Room: Conference I (80)
ICSA Student Paper Awards Committee.
Chair: Ying Hung, Rutgers University.
1:00 PM Optimal calibration for computer model prediction with finite samples
  - Xiaowu Dai. University of Wisconsin-Madison
1:25 PM Empirical-likelihood-based criteria for model selection on marginal analysis of longitudinal data with dropout missingness
  - Chixiang Chen and Ming Wang. Pennsylvania State University College of Medicine
1:50 PM Online Sequential Leveraging Sampling Method for Streaming Time Series Data
  - Ruix. T. N. Sriram\textsuperscript{1}, Wei Biao Wu and Ping Ma\textsuperscript{1}. 1University of Georgia
2:15 PM Floor Discussion.

Session 77: Recent Advances of Statistical Modeling in Biomedical Research (Invited)
Room: Brunswick D (80)
Organizer: Zhao Ren, University of Pittsburgh.
Chair: Zhao Ren, University of Pittsburgh.
1:00 PM Conditional adaptive Bayesian spectral analysis of nonstationary biomedical time series
  - Robert Krafft\textsuperscript{1}, Scott Bruce\textsuperscript{2}, Matrica Hall\textsuperscript{3} and Daniel Buysse\textsuperscript{1}. 1University of Pittsburgh 2George Mason University
1:25 PM BAMM-SC: A Bayesian Mixture Model for Clustering Droplet-based Single cell Transcriptomic Data from multiple individuals
  - Ying Ding\textsuperscript{1}, Zhe Sun\textsuperscript{1}, Ming Ha\textsuperscript{2} and Wei Chen\textsuperscript{1}. 1University of Pittsburgh 2Cleveland Clinic Lerner Research Institute
1:50 PM Understanding Cellular Heterogeneity Using Single-cell Data
  - Lynn Lin. Pennsylvania State University
2:15 PM Fast and Robust Deconvolution of Tumor Infiltrating Lymphocyte from Expression Profiles using Least Trimmed Squares
  - Yuqing Xie\textsuperscript{1}, Ming Yan, Yu Lei and Yuning Hao. 1Michigan State University
2:40 PM Floor Discussion.

June 16 3:00 PM -4:40 PM

Session 78: Statistical Inference in Air Pollution and Health Epidemiology (Invited)
Room: Garden State A (50)
Organizer: Shu Yang, North Carolina State University.
Chair: Shu Yang, North Carolina State University.
3:00 PM A Causal Inference Analysis of the Effect of Wildland Fire Smoke on Ambient Air Pollution Levels
  - Alexandra Larsen\textsuperscript{1}, Shu Yang\textsuperscript{1}, Ana Rappold\textsuperscript{2} and Brian Reich\textsuperscript{1}. 1North Carolina State University 2U.S. Environmental Protection Agency
3:20 PM Efforts to Quantify the Causal Effect of Fine Particulate Matter on Mortality
  - Zhulin He, Zhengyuan Zhu and Hengfang Wang. Iowa State University
3:40 PM Remove Confounding in Air Pollution Studies With Negative Controls
  - Wang Miao and Eric Tchetgen Tchetgen. Harvard University
4:00 PM A Bayesian critical window variable selection method for estimating the impact of air pollution exposure during pregnancy
  - Joshua Warren. Department of Biostatistics, Yale University
4:20 PM Causal inference in the context of an error prone exposure: air pollution and mortality
  - Xiao Wu\textsuperscript{1}, Danielle Braun\textsuperscript{2}, Mariannthi-Anna Kioumourtzoglou\textsuperscript{3}, Christine Choirat\textsuperscript{2}, Qian Di\textsuperscript{2} and Francesca Dominici\textsuperscript{2}. 1Harvard University 2Harvard T.H. Chan School of Public Health 3Mailman School of Public Health, Columbia University
4:40 PM Floor Discussion.
Session 79: New research strategies in clinical studies (Invited)
Room: Conference A (65)
Organizer: Peng Yang, Clindata Insight, Inc.
Chair: Kun Nie, Amgen.
3:00 PM Basket Trial Design in Oncology Studies
  Kun Nie. Clindata Insight, Inc.
3:25 PM Statistical learning for biomarker discovery and development in cancer prevention studies
  Xi Zhou. Weill Cornell Medical College
3:50 PM Penalized multiple inflated values selection method with application to SAFER data
  Yang Li. Renmin University of China
4:15 PM Floor Discussion.

Session 80: Addressing the issue of multiplicity and selection in modern data analysis (Invited)
Room: Regency A (80)
Organizer: Zhigen Zhao, Temple University.
Chair: Zhigen Zhao, Temple University.
3:00 PM Multiple Testing for temporal pattern change detection and identification
  Zhi Wei. New Jersey Institute of Technology
3:25 PM Estimation with truncated data
  ♦Asaf Weinstein1 and Amit Meir. 1Stanford University
3:50 PM Local False Discovery Rate Based Methods for Multiple Testing of One-Way Classified Hypotheses
  ♦Zhigen Zhao1 and Sanat Sarkar. 1Temple University
4:15 PM Floor Discussion.

Session 81: New methods with modern large and complex data sets (Invited)
Room: Brunswick A (120)
Organizer: Cheng Yong Tang, Temple University.
Chair: Zeda Li, Baruch College, The City University of New York.
3:00 PM Pre-processing with Orthogonal Decompositions for High-dimensional Explanatory Variables
  Xu Han. Temple University
3:25 PM Multiple change point detection for manifold-valued data with applications to dynamic functional connectivity
  Qiang Sun. University of Toronto
3:50 PM Empirical Band Analysis of Nonstationary Time Series
  Scott Bruce. George Mason University
4:15 PM A Bayesian General Linear Modeling Approach to Cortical Surface fMRI Data Analysis
  Yu Yue. Baruch College, The City University of New York
4:40 PM Floor Discussion.

Session 82: Recent Advances in Genomic Data Analysis (Invited)
Room: Regency B (80)
Organizer: Mei-Ling Ting Lee, University of Maryland, College Park.
Chair: Weiliang Qiu, Harvard University.
3:00 PM Pedigree-based Functional Linear Mixed Models for Association Analysis of Quantitative Traits with Next-generation Sequencing Data
  •Razong Fan1, Bingsong Zhang1, Ao Yuan1, Chi-Yang Chiu2, Fang Yuan3, Xin Li3, Hong-Bin Fang1, Kenneth Lange4, Daniel Weeks5, Alexander Wilson6, Joan Bailey-Wilson6, M’hamed Lajmil-Chaieb7, Richard Cook8, Francis McMahan9, Christopher Amos9 and Miao Xiong10. 1Georgetown University Medical Center
2University of Tennessee Health Science Center 3Kunming Medical University 4University of California, Los Angeles 5University of Pittsburgh 6National Institutes of Health 7Université Laval 8University of Waterloo 9Dartmouth Medical School 10University of Texas at Houston
3:25 PM Gene-based Association Testing of Dichotomous Traits with Generalized Linear Mixed Models Using Extended Pedigrees
  •Chi-Yang Chiu1, Yingda Jiang, Daniel Weeks5 and Razong Fan1. 1The University of Tennessee Health Science Center 2University of Pittsburgh 3Georgetown University Medical Center
3:50 PM Meta-analytic and integrative framework for sparse K-means to identify disease subtypes
  •Zhiguang Huo1 and George Tseng2. 1Department of Biostatistics, University of Florida 2University of Pittsburgh
4:15 PM RNASeqDesign: A framework for RNA-Seq genome-wide power calculation and study design issues
  •Chien-Wei Lin1, George Tseng2, Yongseok Park2, Ge Liao and Mei-Ling Ting Lee3. 1Medical College of Wisconsin 2University of Pittsburgh 3University of Maryland, College Park
4:40 PM Floor Discussion.

Session 83: Statistical Learning for Complex and High-Dimensional Data with Modern Applications (Invited)
Room: Garden State B (70)
Organizer: Lingzhou Xue, Pennsylvania State University.
Chair: Lingzhou Xue, Pennsylvania State University.
3:00 PM Blessing of Massive Scale: Total Cardinality Approach for Spatial Graphical Model Estimation
  Ethan Fang. Pennsylvania State University
3:25 PM Semi-supervised Inference for Explained Variance in High-dimensional Linear Regression and Its Applications
  •Zijian Guo1 and Tony Cai2. 1Rutgers University 2University of Pennsylvania
3:50 PM Learning causal networks via additive faithfulness
  •Kuang-Yao Lee1, Tianqi Liu2, Bing Li3 and Hongyu Zhao4. 1Temple University 2Kensho 3Pennsylvania State University 4Yale University
4:15 PM Minimax Estimation of Large Precision Matrices with Bandable Cholesky Factor
  Zhao Ren. University of Pittsburgh
4:40 PM Floor Discussion.
Session 84: Recent advances in high-dimensional statistics
(Invited)
Room: Brunswick B (120)
Organizer: Zijian Guo, Rutgers University.
Chair: Shaokun Li, University of Pennsylvania.
3:00 PM Adaptive QDA to dimensionality
Qi Yang1 and Guang Cheng2.
1Department of Statistics, Purdue University 2Purdue University
3:25 PM Robust estimation, efficiency, and Lasso debiasing
Po-Ling Loh. University of Wisconsin-Madison
3:50 PM Learning Nonconvex Hierarchical Interactions
Lingzhou Xue1 and Hongbo Dong. Pennsylvania State University
4:15 PM Randomized incomplete U-statistics in high dimensions
Xiaohui Chen1 and Kengo Kato2.
1University of Illinois at Urbana–Champaign 2University of Tokyo
4:40 PM Floor Discussion.

Session 85: Nonparametric and high dimensional statistics
(Invited)
Room: Garden State C (70)
Organizer: Zongming Ma, University of Pennsylvania.
Chair: Fengnan Gao, Fudan University.
3:00 PM Mean Field Variational Inference: Computational and Statistical Guarantees
Anderson Zhang. Yale University
3:25 PM Trace regression with nonconvex regularization
Tingni Sun. University of Maryland, College Park
3:50 PM Maximum likelihood estimation of Sublinear Preferential Attachment Models and its connection to Janson’s urn models
Fengnan Gao1 and Aad Van der Vaart. Fudan University
4:15 PM Floor Discussion.

Session 86: Leveraging Evidence from Historical/Subsidiary Data in Randomized Controlled Trial (Invited)
Room: Conference B (65)
Organizer: Samiran Ghosh, Wayne State University.
Chair: Freda Cooner, Sanofi.
3:00 PM The use of alternative data in clinical trials in small population and pediatric patients
Shiling Ruan. Novartis
3:25 PM Bayesian Approach of testing Noninferiority in Presence of Historical Data
Samiran Ghosh. School of Medicine, Wayne State University
3:50 PM A Note on Posterior Predictive Assessment to Assess Model Fit for Incomplete Longitudinal Data
Arkendu Chatterjee. Merck & Co., Inc.
4:15 PM Discussant: Yu Lan, Sanofi and Satrajit Roychoudhury, Pfizer Inc.
4:40 PM Floor Discussion.

Session 87: Immunogenicity Testing of Therapeutic Protein Product (Invited)
Room: Conference C (65)
Organizer: Yi Tsong, CDER, FDA.
Chair: Shein-Chung Chow, CDER, FDA.
3:00 PM Issues in designing immunogenicity studies
Yun Wang. U.S. Food and Drug Administration
3:25 PM Statistical evaluation of several methods for cut point determination of immunogenicity screening assay
Yi Tsong1, Mei Yu Shen2 and Xiaoyu Dong3.
1CDER, FDA 2U.S. Food and Drug Administration 3Amgen
3:50 PM Discussant: Shein-Chung Chow, CDER, FDA.
4:15 PM Floor Discussion.

Session 88: Statistical advances and challenges in immuno-oncology development (Invited)
Room: Conference I (80)
Organizer: Qi Qi Deng, Boehringer Ingelheim Pharmaceuticals, Inc.
Chair: Wenqiong Xue, Boehringer Ingelheim Pharmaceuticals, Inc.
3:00 PM Robust Phase 3 Designs for Adaptive Population Modification
Cong Chen1, Xiaoyun(Nicole) Li, Wen Li and Robert Beckman. 1Merck & Co., Inc.
3:25 PM A Bayesian hierarchical model for indirect comparison of immuno-oncology drugs
Jingyi Liu, Ji Lin, Zach Thomas and Yumin Zhao. Eli Lilly and Company
3:50 PM Statistical Considerations When Applying Restricted Mean Survival Time in Immune-Oncology Trials
Ying Lu1, Hua Jin2 and Lu Tian1.
1Department of Biomedical Data Science and Center for Innovative Study Design, Stanford University 2South China Normal University and Stanford University
4:15 PM Discussant: Qi Qi Deng, Boehringer Ingelheim Pharmaceuticals, Inc.
4:40 PM Floor Discussion.

Session 89: Statistical learning for high-dimensional and complex data (Invited)
Room: Brunswick C (80)
Organizer: Yiwen Liu, University of Georgia.
Chair: Wenzhang Zhong, University of Georgia.
3:00 PM Theory Informs Practice: Smoothing Parameters Selection for Smoothing Spline ANOVA Models in Large Samples
Xiao Xia Sun1, Wenzhang Zhong2 and Ping Ma2.
1University of Arizona 2University of Georgia
3:25 PM High-dimensional Spatially Varying Coefficient Models
Xinyi Li1, Li Wang1, Dan Nettleton2 and Guannan Wang2.
1Iowa State University 2College of William & Mary
3:50 PM Differential network analysis via a novel biclustering framework
Sha Cao1, Xiaoyu Lu, Chi Zhang1 and Na Bo2.
1Indiana University School of Medicine
June 17 8:30 AM - 10:10 AM

Session 90: Advances in Complex and Network Data Analysis (Invited)
Room: Regency C (80)
Organizer: Pengsheng Ji, University of Georgia.
Chair: Pengsheng Ji, University of Georgia.

3:00 PM On the generalization and computation of data depth
Yiyuan She. Florida State University

3:25 PM Two-Step Bayesian Multiple Classifications with Logic Expressions
♦ Wensong Wu and Tan Li. Florida International University

3:50 PM Understanding Kernel-Embedding Based Goodness-of-Fit Tests
Krishnakumar Balasubramanian. Princeton University

4:15 PM An empirical bayes method to estimate interaction intensities and identify long-range chromosomal interactions based on Hi-C data
♦ Qi Zhang and Zheng Xu. University of Nebraska–Lincoln

4:40 PM Floor Discussion.

Session 101: The win ratio: what is it? (Invited)
Room: Brunswick D (80)
Organizer: Gaohong Dong, iStats, Inc.
Chair: Haiyan Xu, Johnson & Johnson; Yaping Wang, FDA.

3:00 PM What is really the win ratio?
♦ Gaohong Dong 1, Victoria Chang 2, Junshan Qiu 3, Roland Matsouaka 4 and David Hoaglin 5. 1iStats, Inc. 2Abbvie, Inc. 3U.S. Food and Drug Administration 4Duke University School of Medicine 5University of Massachusetts Medical School

3:25 PM A Generalized Analytic Solution to the Win Ratio
♦ Di Li 1 and Gaohong Dong 2. 1Bristol-Myers Squibb 2iStats, Inc.

3:50 PM Confidence interval estimations for the matched win ratio
♦ Roland Matsouaka 1 and Adrian Coles 2. 1Duke University School of Medicine 2Duke University

4:15 PM Stratified win ratio
♦ Junshan Qiu 1, Gaohong Dong 2, Duolao Wang and Marc Vandemeulebroecke. 1U.S. Food and Drug Administration 2iStats, Inc.

4:40 PM Discussant: Xiaodong Luo, Sanofi
Floor Discussion.
8:30 AM Partitioned Weighted Approach for Estimating Marginal Posterior Density with Applications
   • Lynn Kuo, Yu-Bo Wang, Ming-Hui Chen and Paul Lewis. University of Connecticut

8:55 AM A Bayesian hierarchical model for related densities using Polya trees
   • Li Ma and Jonathan Christensen. Duke University

9:20 AM Semi-Implicit Variational Inference
   • Mingyuan Zhou and Mingzhang Yin. University of Texas at Austin

9:45 AM Bayesian Integrative Model for Deciphering High-Dimensional Genotype-Phenotype Map
   Chuanhua Julia Xing. XPrecision, LLC

10:10 AM Floor Discussion.

Session 94: Statistical methods for large-scale complex data (Invited)
Room: Garden State A (50)
Organizer: Xin Xing, University of Georgia.
Chair: Wenxuan Zhong, University of Georgia.

8:30 AM Integrative statistical causal analysis of the microbiome, metabolome and inflammation data in to understand oral cancer mechanism
   • Di Wu, Siliang Gong, Ruijin Wang and Flavia Teles. University of North Carolina at Chapel Hill University of Pennsylvania

8:55 AM Pearson’s chi-squared statistics: approximation theory and beyond
   • Mengyu Xu, Danna Zhang and Weibiao Wu. University of Central Florida

9:20 AM Early stopping for nonparametric testing
   • Meimei Liu and Guang Cheng. Purdue University

9:45 AM Left truncated mixture Gaussian distribution based modeling of single cell RNA-seq data
   • Chi Zhang and Changlin Wan. Indiana University School of Medicine

10:10 AM Floor Discussion.

Session 95: Statistical Methods Beyond Parametrics in Biomedical Studies (Invited)
Room: Regency C (80)
Organizer: Jiwei Zhao, State University of New York at Buffalo.
Chair: Jiwei Zhao, State University of New York at Buffalo.

8:30 AM Nonparametric Identified Methods to Handle Nonignorable Missing Data
   Mauricio Sadinle. University of Washington

8:55 AM Recurrent Event Analysis with Covariates Observed at Irregular, Informative Clinical Visits, with Application to Electronic Medical Records Data
   • Yifei Sun and Chiang-Yu Huang. Columbia University University of California, San Francisco

9:20 AM Some Ideas on Calibration in the Presence of Nonignorable Missing Data
   Peisong Han. University of Michigan

9:45 AM Mediation Higher Criticism Test in Epigenomic Studies
   • Jincheng Shen1 and Xihong Lin2. University of Utah

10:10 AM Floor Discussion.

Session 96: High-dimensional statistical inference and computing (Invited)
Room: Brunswick A (120)
Organizer: Anru Zhang, University of Wisconsin-Madison.
Chair: Anru Zhang, University of Wisconsin-Madison.

8:30 AM NEST: Nonparametric empirical Bayes smoothing Tweedie’s estimator under heteroscedasticity
   Wenguang Sun. University of Southern California

8:55 AM Random Initialization and Implicit Regularization in Non-convex Phase Retrieval
   • Yuxin Chen1, Cong Ma1, Yuejie Chi2 and Jiaying Fan1. Princeton University Carnegie Mellon University

9:20 AM Tensor SVD: Statistical and Computational Limits
   Anru Zhang. University of Wisconsin-Madison

9:45 AM Teaching and learning in uncertainty
   Varun Jog. University of Wisconsin-Madison

10:10 AM Floor Discussion.

Session 97: Recent advances in statistical analysis of high-dimensional biological data (Invited)
Room: Conference C (65)
Organizer: Tao He, Department of Mathematics, San Francisco State University.
Chair: Yu Jiang, University of Memphis.

8:30 AM A semi-supervised approach for predicting cell type/tissue specific functional consequences of non-coding variation using massively parallel reporter assays
   • Zihua He1, Linxi Liu, Kai Wang and Juliana Ionita-Laza. Columbia University

8:55 AM Clustering of Functional Data
   • Sneha Jadhav and Shuangge Ma. Yale University

9:20 AM STEPS: An Efficient Prospective Likelihood Approach to Genetic Association analyses of Secondary Traits in Extreme Phenotype
   • Guolian Kang2, Yun Li, Matthew Smeltzer, Guimin Gao and Shengli Zhao. St. Jude Children’s Research Hospital

9:45 AM Dissecting the Gene-environment Interactions via Bayesian Hierarchical models
   • Yu Jiang1 and Cen Wu2. University of Memphis Kansas State University

10:10 AM Floor Discussion.

Session 98: Inference for large complex data (Invited)
Room: Conference I (80)
Organizer: Xiangyu Wang, Google, LLC.
Chair: Binyan Jiang, The Hong Kong Polytechnic University.

8:30 AM Floor Discussion.
Session 99: Data Science Intersecting with Health Policy
(Invited)
Room: Brunswick B (120)
Organizer: Kelly Zou, Pfizer Inc.
Chair: Shin Ah Oh, Pfizer Inc.
8:30 AM Composite Interaction Tree for Simultaneous Learning Optimal Individualized Treatment Rules and Subgroups
Xin Qiu and Yuanjia Wang. Columbia University
8:55 AM Text Mining of Policy Documents
Shin Ah Oh, Kelly Zou, Jason Pan, Ching-Ray Yu and Martin Carlsson. Pfizer Inc.
9:20 AM The future is now, but where we should focus
Haoda Fu. Eli Lilly and Company
9:45 AM Floor Discussion.

Session 100: Causal Inference and Covariate Balance (Invited)
Room: Garden State B (70)
Organizer: Yeying Zhu, Department of Statistics & Actuarial Science, University of Waterloo.
Chair: Yeying Zhu, Department of Statistics & Actuarial Science.
8:30 AM Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores
Shu Yang. North Carolina State University
8:55 AM Propensity score weighting analysis and treatment effect discovery
Liang Li. MD Anderson Cancer Center
9:20 AM Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data
Zhiqiang Tan. Rutgers University
9:45 AM Assessing covariate balance for clustered data with continuous exposures
Donna Coffman1 and Megan Schuler2. 1Temple University 2RAND Corp
10:10 AM Floor Discussion.

Session 102: Contributed Session 1 (Contributed)
Room: Conference A (65)
Chair: Dipankar Bandyopadhyay, Virginia Commonwealth University.
8:30 AM Intensity Estimation of Spatial Point Processes Based on Area-Aggregated Data
Chi-Wei Lai and Hsin-Cheng Huang. Institute of Statistical Science, Academia Sinica
8:50 AM Bayesian Nonparametric Policy Search with Application to Periodontal Recall Intervals
Dipankar Bandyopadhyay1, Brian Reich2, Eric Laber2 and Qian Guan2. 1Virginia Commonwealth University 2North Carolina State University
9:10 AM Quadratic Discriminant Analysis by Projection
Ruiyang Wu and Ning Hao. University of Arizona
9:30 AM On asymptotic risk of order selection in integrated autoregressive models
Shu-Hui Yu. Institute of Statistics, National University of Kaohsiung
9:50 AM A Principal Stratification Approach to Evaluate the Causal Effect of A Patient Activation Intervention For Bone Health Outcomes
Yiyue Luo1, Michael Jones2, Fredric Wolinsky2, Peter Cram3 and Stephanie Edmonds4. 1Vertex Pharmaceuticals 2University of Iowa College of Public Health 3Toronto General Hospital Research Institute (TGHRI) 4Woman Centered Health

Session 103: Contributed Session 2 (Contributed)
Room: Conference B (65)
Chair: Rongning Wu, City University of New York.
8:30 AM Prediction Interval for Autoregressive Time Series via Oracally Efficient Estimation of Multi-Step Ahead Innovation Distribution Function
Lijie Gu1, Juanjuan Kong and Lijian Yang2. 1Soochow University 2Tsinghua University
8:50 AM Assessing covariate balance for clustered data with continuous exposures
Donna Coffman and Megan Schuler. 1Temple University 2RAND Corp
9:10 AM Least tail-trimmed absolute deviation estimation for autoregressions with infinite/finite variance
Rongning Wu1 and Yunwei Cui2. 1City University of New York 2Towson University
9:30 AM A Principal Stratification Approach to Evaluate the Causal Effect of A Patient Activation Intervention For Bone Health Outcomes
Runmin Wang and Xiaofeng Shao. University of Illinois at Urbana–Champaign

10:10 AM Floor Discussion.
June 17 10:30 AM -12:10 PM

Session 104: Real-time prediction of clinical trial enrollment and event times (Invited)
Room: Garden State B (70)
Organizer: Yu Lan, Sanofi.
Chair: Shuang (Nancy) Li, Southern Methodist University.
10:30 AM Blinded sample size re-estimation in trials with survival outcomes and incomplete information.
Thomas Cook. University of Wisconsin-Madison
10:55 AM Real-time prediction of event times in clinical trials - an overview
Gui-Shuang Ying. University of Pennsylvania
11:20 AM Adaptive parametric prediction of event times in clinical trials
*Lan Yu¹ and Daniel Heitjan². ¹Sanofi ²Southern Methodist University & University of Texas Southwestern Medical Center
11:45 AM Bayesian Modeling and Prediction of Patient Accrual in Multicenter Clinical Trials with Varying Center Activation Times
*Junhao Liu¹, Jo Wick¹, Yu Jiang², Matthew Mayo² and Byron Gajewski¹. ¹University of Kansas Medical Center ²University of Memphis
12:10 PM Floor Discussion.

Session 105: Analysis of Data from Wearable Devices (Invited)
Room: Regency C (80)
Organizer: Robert Krafty, University of Pittsburgh.
Chair: Robert Krafty, University of Pittsburgh.
10:25 AM Robust functional principal components analysis with application to accelerometer data
*Sisheng Liu¹, Chongzhi Di¹ and Fang Han². ¹Fred Hutchinson Cancer Research Center ²University of Washington
10:55 AM Multilevel variance components model in functional data with application in minute-level accelerometry measures
Haochang Shou. University of Pennsylvania
11:20 AM Gait characteristics extracted from raw accelerometry data collected in free-living settings
Jacek Urbanek. Johns Hopkins School of Medicine
11:45 AM Discussant: Donna Coffman, Temple University
12:10 PM Floor Discussion.

Session 106: New developments in large-scale data analysis (Invited)
Room: Regency B (80)
Organizer: Ping-Shou Zhong, Michigan State University.
Chair: Shawn Santo, Michigan State University.
10:30 AM Sorted Concave Penalized Regression
*Long Feng³ and Cun-Hui Zhang. ¹Yale University
10:55 AM Deep Learning Approaches for Agricultural Plant Phenotyping Based on RGB Images and Hyperspectral Images
*Zheng Xu, James Schnable and Jennifer Clarke. University of Nebraska–Lincoln
11:20 AM Homogeneity tests of covariance matrices with high-dimensional longitudinal data
*Shawn Santob¹, Ping-Shou Zhong¹ and Runze Li². ¹Michigan State University ²Pennsylvania State University
11:45 AM Floor Discussion.

Session 107: Machine Learning for Healthcare Applications (Invited)
Room: Garden State C (70)
Organizer: Ping Wang, Insmed.
Chair: Hongwei Wang, Abbvie, Inc.
10:30 AM Traditional Statistics and Machine Learning using Electronic Health Records Data
*Yirui Hu and H. Lester Kirchner. Geisinger
10:55 AM Clinical NLP and deep learning for disease classification and reporting in Chest X-rays
Yifan Peng. National Institutes of Health
11:20 AM Transformation in Real World Evidence through Machine Learning
Jyotsna Mehta. founder of KEVA health
11:45 AM Discussant: Hongwei Wang, Abbvie, Inc.
12:10 PM Floor Discussion.

Session 108: Reasonable Possibility is Statistical Science in Drug Safety Monitoring (Invited)
Room: Conference I (80)
Organizer: LiAn Lin, Merck & Co., Inc.
Chair: LiAn Lin, Merck & Co., Inc.
10:30 AM Reasonable Possibility in the Pre- vs Post-Approval Settings
Ed Whalen. Pfizer Inc.
10:55 AM Blinded vs Unblinded Analysis for IND safety reporting
Lian Lin. Merck & Co., Inc.
11:20 AM Visual Analytics Tools to Assess Possible Causality of a Drug-Induced Adverse Event of Special Interest
*Krishan Singh¹, Melvin Munsaka² and Kefei Zhou³. ¹GlaxoSmithKline ²Abbvie, Inc. ³Theravance
11:45 AM Meta-Analysis Techniques and Safety Monitoring
Rositsa Dimova. U.S. Food and Drug Administration
12:10 PM Discussant: Rositsa Dimova, U.S. Food and Drug Administration and William Wang, Merck & Co., Inc.
Floor Discussion.

Session 109: Integrative Statistical Learning in High Dimensional Omics Data Analysis (Invited)
Room: Brunswick C (80)
Organizer: Yu Jiang, University of Memphis.
Chair: Cen Wu, Kansas State University.
10:30 AM Integrative Gene-environment Interactions From Multidimensional Omics Data In Cancer Prognosis
*Cen Wu and Yinshao Du. Kansas State University

2018 ICSA Applied Statistics Symposium, New Brunswick, NJ, June 14-17
10:55 AM Integrated association testing of multiple forms of omics data with an endpoint of interest
• Xueyuan Cao¹, Mingjuan Wang², Dale Bowman³, Stanley Pounds³ and E. Olusegun George³. ¹University of Tennessee Health Science Center ²St. Jude Children’s Research Hospital ³University of Memphis

11:20 AM A random-effects model for multi-tissue deconvolution to estimate individual-level cell-type-specific gene expression
• Jiebiao Wang¹, Bernie Devlin² and Kathryn Roeder¹. ¹Carnegie Mellon University ²University of Pittsburgh

11:45 AM Intensity normalization of MR images across subjects as a pre-processing tool to improve analysis and segmentation results of MS drawn from studying unwanted variation effects in gene expressions
• Abhita Amitabha Sarkar and Russell Shinhara. University of Pennsylvania

12:10 PM Floor Discussion.

Session 110: Recent advances in machine learning and causal inference (Invited)
Room: Regency A (80)
Organizer: Yang Ning, Cornell University.
Chair: Yang Ning, Cornell University.

10:30 AM Distance-based and RKHS-based Dependence Metrics in High-dimension
• Changbo Zhu¹, Shun Yao², Xianyang Zhang³ and Xiaofeng Shao¹. ¹University of Illinois at Urbana–Champaign ²Goldman Sachs, New York City ³Texas A & M University

10:50 AM Pathway and Gene Selection with Guided Regularized Random Forests
• Tyler Cook¹, Daniel Brunley¹ and Sounak Chakraborty². ¹University of Central Oklahoma ²University of Missouri

11:45 AM Discussant: Yi Tsong, U.S. Food and Drug Administration.

Session 111: Statistical Considerations of In-Vitro Bioequivalence or Adhesion Data Analysis (Invited)
Room: Brunswick D (80)
Organizer: Yu-Ting Weng, U.S. Food and Drug Administration.
Chair: Mengdie Yuan, U.S. Food and Drug Administration.

10:30 AM In vitro Permeation Testing- A scientific overview
• Sarah Ibrahim. U.S. Food and Drug Administration

10:55 AM Application of scaled bioequivalence methods to in vitro permeation testing experiments
• Kimberly Walters and Lisa Weisfeld. Statistics Collaborative, Inc.

11:20 AM Application of mixed models to in vitro permeation testing experiments for the assessment of dose level in transdermal patches
• Lisa Weisfeld and Kimberly Walters. Statistics Collaborative, Inc.
Scientific Program (\textit{Presenting Author})

11:10 AM Semiparametric Dynamic Adaptive Robust Estimations for Varying High-Dimensional Networks
\textit{Tzu-Chun Wu and Emily Kang}. University of Cincinnati

11:30 AM Gaussian Sparse Partial Membership Model with Applications to Cancer Genomics
\textit{Wei Xia, Clint Pazhayidam George and George Michailidis}. University of Florida

11:50 AM A Randomized Algorithm for Maximum Likelihood Estimation with Spatial Autoregressive Models Applied on Social Network Data
\textit{Miaoqi Li and Emily Kang}. University of Cincinnati

12:10 PM Floor Discussion.

\textbf{Session 115: Contributed Session 6 (Contributed)}

Room: Conference C (65)
Chair: Liang Zhu, UT Health.

10:30 AM Variable selection of lipid-environment interactions in longitudinal studies
\textsuperscript{1}Kansas State University \textsuperscript{2}University of Memphis

10:50 AM Statistical Analysis on Clustered Mixed Recurrent Event data
\textit{Liang Zhu}\textsuperscript{1}, \textit{Yimei Li}\textsuperscript{2} and \textit{Leslie Robison}\textsuperscript{2}.  
\textsuperscript{1}UT Health \textsuperscript{2}St. Jude Children’s Research Hospital

11:10 AM On comparing 2 correlated C indices with censored survival data
\textit{Yilong Zhang}, Xiaoxia Han and Yongzhao Shao.  
\textsuperscript{1}Merck Research Laboratories \textsuperscript{2}Public Health Sciences Department, Henry Ford Health System \textsuperscript{3}Department of Population Health, New York University School of Medicine.

11:30 AM Avoiding Bias in Internal Pilot Design for Balanced Repeated Measures
\textit{Xinrui Zhang} and Yueh-Yun Chi.  
\textsuperscript{1}Novartis \textsuperscript{2}University of Florida

11:50 AM Analysis of Generalized Semiparametric Varying-Coefficient Effects Models for Longitudinal Data
\textit{Li Qi} and Yanqing Sun.  
\textsuperscript{1}Sanofi \textsuperscript{2}University of North Carolina at Chapel Hill

12:10 PM Floor Discussion.
Keynote lecture I

Keynote lecture I: Reflections on Statistical Practice, Leadership, and Training
Lisa LaVange
University of North Carolina at Chapel Hill

Having recently completed six years in a statistical leadership role at FDA, I will reflect on the importance and impact of statistics in drug development and regulation, citing examples from policy development and dissemination in addition to statistical reviews of new drugs. Opportunities for FDA statisticians to learn about and practice leadership in their work environment will be compared and contrasted with opportunities in other employment sectors. Training statisticians to lead, formally and informally, is essential for the future of our profession, and I will offer my view for ways this can be accomplished, including highlighting the planned initiatives of the ASA Statistical Leadership Institute.

Session 1: Network Data Analysis

Model-assisted design of experiments on social networks
Eduardo Airoldi
Harvard University
airolidi@fas.harvard.edu

Classical approaches to causal inference largely rely on the assumption of "lack of interference", according to which the outcome of an individual does not depend on the treatment assigned to others, as well as on many other simplifying assumptions, including the absence of strategic behavior. In many applications, however, such as evaluating the effectiveness of healthcare interventions that leverage social structure, or assessing the impact of product innovations and ad campaigns on social media platforms, or experimentation at scale in large IT companies, assuming lack of interference and other simplifying assumptions is untenable. Moreover, the effect of interference itself is often an inferential target of interest, rather than a nuisance. In this talk, we will formalize technical issues that arise in estimating causal effects when interference can be attributed to a network among the units of analysis, within the potential outcomes framework. We will introduce and discuss several strategies for experimental design in this context centered around a judicious use of statistical models, which we refer to as "model-assisted" design of experiments. In particular, we wish for certain finite-sample properties of the estimator to hold even if the model catastrophically fails, while we would like to gain efficiency if certain aspects of the model are correct. We will then contrast design-based, model-based and model-assisted approaches to experimental design from a decision theoretic perspective.

Why Aren’t Network Statistics Accompanied By Uncertainty Statements?
Eric Kolaczyk
Boston University
kolaczyk@bu.edu

Over 500K articles have been published since 1999 with the word "network" in the title. And the vast majority of these report network summary statistics of one type or another. However, these numbers are rarely accompanied by any quantification of uncertainty. Yet any error inherent in the measurements underlying the construction of the network, or in the network construction procedure itself, necessarily must propagate to any summary statistics reported. Perhaps surprisingly, there is little in the way of statistical methodology for this problem. I summarize results from our recent work, for the case of estimating the density of low-order subgraphs. Under a simple model of network error, we show that consistent estimation of such densities is impossible when the rates of error are unknown and only a single network is observed. We then develop method-of-moment estimators of subgraph density and error rates for the case where a minimal number of network replicates are available. These estimators are shown to be asymptotically normal as the number of vertices increases to infinity. We also provide confidence intervals for quantifying the uncertainty in these estimates based on the asymptotic normality. We illustrate the use of our estimators in the context of gene coexpression networks. This is joint work with Qiwei Yao and Jinyuan Chang.

Empirical risk minimization for relational data
Peter Orbanz
Columbia University
po2197@columbia.edu

Consider a prediction problem involving relational data—roughly, any data set that can be represented as a large graph or matrix. Such problems include node classification, link prediction, user preference prediction, etc. Optimal predictors for other types of data, where an i.i.d. assumption can be justified, are usually chosen using empirical risk minimization, which in turn is often combined with stochastic gradient descent algorithms. Extending these ideas to graph data has proven difficult. I will explain how one can (i) define an empirical risk for relational data and (ii) obtain stochastic gradients for this risk that are automatically unbiased. The key ingredient is to consider the method by which data is sampled from a graph as an explicit component of model design. This is a joint work with Victor Veitch, Wenda Zhou, Morgane Austern and David Blei.

Session 2: Future of Statisticians in the pharmaceutical world – A panel Discussion with Leaders

Panel discussions.

Session 3: Some recent development in large-scale statistical learning with complex data

Integrative multi-view reduced-rank regression: bridging group-sparse and low-rank models.
*Gen Li¹ and Kun Chen²
¹Columbia University
²University of Connecticut
gj2521@columbia.edu

In this talk, I will introduce a novel integrated reduced-rank framework for multivariate regression. Predictors are multi-view data, which naturally form different groups. Each predictor group has its
unique low-rank coefficient matrix. The framework flexibly captures the relationship between multivariate responses and predictors, and subsumes many existing methods such as reduced rank regression and group lasso as special cases. We develop an efficient alternating direction method of multipliers (ADMM) algorithm for model fitting, and exploit a majorization approach to deal with binary responses or missing values in responses. We demonstrate the efficacy of the proposed methods with simulation studies and a real application to the Longitudinal Study of Aging.

Sparse Penalized Quantile Regression: Method, Theory, and Algorithm
Yiwen Gu
University of Connecticut

Sparse penalized quantile regression is a useful tool for variable selection, robust estimation, and heteroscedasticity detection in high-dimensional data analysis. We discuss the variable selection and estimation properties of the lasso and folded concave penalized quantile regression via non-asymptotic arguments. We also consider consistent parameter tuning therein. The computational issue of the sparse penalized quantile regression has not yet been fully resolved in the literature, due to non-smoothness of the quantile regression loss function. We introduce fast alternating direction method of multipliers (ADMM) algorithms for computing the sparse penalized quantile regression. Numerical examples demonstrate the competitive performance of our algorithm: it significantly outperforms several other fast solvers for high-dimensional penalized quantile regression.

Disease progression modeling with large-scale observational data
Zhaonan Sun
IBM Research

Chronic diseases such as COPD, Diabetes, and Huntington’s Disease progress slowly over an extended period, causing severe emotional and financial distress to patients, their families, and the healthcare system at large. A better understanding of the progression of such diseases is instrumental in early diagnosis and personalized care. However, understanding disease progression from real-world evidence is challenging. Not only are observations noisy, and irregular in time, but the rate of progression may exhibit significant variance across patients. The knowledge of how progression is affected by differing patient characteristics is crucial for both better understanding the biological processes underlying the disease and for informing clinical decisions such as the patient recruitment and trial design. We propose a statistical disease progression model that explicitly accounts for patient level heterogeneity in progression by conditioning on patient characteristics. Further, the model is able to learn a continuous-time progression from discrete-time observations with non-equal intervals, and stitch together an entire progression trajectory from a set of incomplete records, each of which only covers short segments of progression. We demonstrate the capabilities of the proposed model by both simulation studies and applying it to a real-world Huntington’s Disease patient cohort and deriving interesting clinical insights.

Efficient big data model selection with applications to fraud detection
Gregory Vaughan
Bentley University

As the volume and complexity of data continues to grow, more attention is being focused on solving so-called big data problems. One field where this focus is pertinent is credit card fraud detection. Model selection approaches can identify key predictors for preventing fraud. Stagewise Selection is a classic model selection technique that has experienced a revitalized interest due to its computational simplicity and flexibility. Over a sequence of simple learning steps, stagewise techniques build a sequence of candidate models that is less greedy than the stepwise approach. This presentation introduces a new stochastic stagewise technique that integrates a subsampling approach into the stagewise framework, yielding a simple tool for model selection when working with big data. Simulation studies demonstrate the proposed technique offers a reasonable trade-off between computational cost and predictive performance. We apply the proposed approach to synthetic credit card fraud data to demonstrate the technique’s application.

Session 4: SIBS Special Session: Methods for Big Data Analysis in Genomics and Biomedicine

Uncovering the hierarchical conformation of topologically associating domains from Hi-C data
Yu Zhang1, Lin An, Tao Yang and Qunhua Li1
1Pennsylvania State University

Motivation: Mammalian genomes are organized into different levels. Observations from chromosome conformation capture methods (Hi-C) suggests chromatin forms frequent local interactions in certain regions, which is called as Topologically Associating Domains (TADs). While TADs are often used as the smallest structure units to study regulatory mechanism, previous observation shows that hierarchy is present in TADs, with smaller TADs nested within larger ones. Though several different TAD calling algorithms have been developed, limited research has been done to reliably identify hierarchical TAD structures and understand their roles in gene regulation.

Result: We developed a new method to call TADs in hierarchy. Our systematic validation based on epigenomics information and gene expression information shows that our method greatly outperforms existing TAD callers in both terms of accuracy and reproducibility. Using this method, we uncovered novel biological insights related to TAD hierarchies. For example, we found the cohesion density has significant positive association with interaction strength within TADs. We found that hierarchical TADs and TADs without hierarchy have many distinct features. Notably, we found that active epigenetic states (promoter associated and enhancer associated) are more enriched in nested TADs than larger TADs. Moreover, our results suggest boundaries in nested TADs tend to be more CTCF-enriched than those in TADs without hierarchical structures. Finally, our results also are in good agreement with the TAD extrusion model.

Conclusion: The new hierarchical TAD caller is able to identify different levels of TADs. Combined with epigenetics and transcriptome information, our results generate new insights towards understanding the complex system of gene regulation.

Estimating the overall contribution of human microbiome to the risk of developing cancers based on prospective studies
Jianxin Shi
National Cancer Institute

2018 ICSA Applied Statistics Symposium, New Brunswick, NJ, June 14-17
The human microbiome, also called as the second human genome, is the collection of microbes inhabiting the human body, including bacteria, archaea, fungi and viruses. Advances in high-throughput sequencing allow characterization of the compositions of human microbial communities and make it possible to perform large-scale epidemiological studies. A large scale prospective microbiome study, collecting microbiome sample and following up for many years, has the potential to identify microbiome risk factors for complex diseases and improve risk prediction. One key issue is to estimate the overall contribution of microbiome to the risk of developing a specific disease. We develop a generalized linear mixed model to estimate such overall contribution, defined as the fraction of phenotypic variance (log(time-to-event)) explained by all OTUs, which is solved by a stochastic approximation algorithm. Preliminary results will be discussed for three oral microbiome studies for multiple-site cancers based on case-cohort design performed at the Division of Cancer Epidemiology and Genetics of National Cancer Institute.

A geometric comparison of the rejection regions of association tests

Ian Barnett1, Zhonghua Liu2 and Xihong Lin2
1University of Pennsylvania
2Harvard University

Whether it is multiple phenotypes or multiple SNPs within a gene that are analyzed jointly, the problem of simultaneously evaluating multiple marginal test statistics is a persistent challenge in genomics. There are many statistical association tests that have been developed for this setting, such as variance component tests, max tests, and the higher criticism. When comparing the power of these methods, there is no universally most powerful method, and the best choice of method is highly dependent on the strength, direction, and sparsity of the signal. In order to gain an intuition behind the strengths and weaknesses of these popular approaches, we evaluate the acceptance/rejection regions of each test to see, geometrically, when they are similar or different from one another. We also explore how different alternatives effect these similarities and differences. We find that what drives the relative test performances is not the sparsity of the signal vector, but rather its direction.

A semi-supervised approach for predicting cell type/tissue specific functional consequences of non-coding variation using massively parallel reporter assays

Iuliana Ionita-Laza
Columbia University

Predicting the functional consequences of genetic variants is a challenging problem, especially for variants in non-coding regions. Projects such as ENCODE and Roadmap Epigenomics make available various epigenetic features, including histone modifications and chromatin accessibility, genome-wide in over a hundred different tissues and cell types. In addition, recent developments in high-throughput assays to assess the functional impact of variants in regulatory regions (e.g. massively parallel reporter assays, CRISPR/Cas9-mediated in situ saturating mutagenesis) can lead to the generation of high quality data on the functional effects of selected variants. We propose here a semi-supervised approach, GenoNet, to jointly utilize experimentally confirmed regulatory variants (labeled variants), millions of unlabeled variants genomewide, and more than a thousand cell type/tissue specific epigenetic annotations to predict functional consequences of non-coding genetic variants. Through the application to several experimental datasets, including massively parallel reporter assay validated variants, and sets of eQTLs and dsQTLs, we demonstrate that the proposed method significantly improves prediction accuracy compared to existing functional prediction methods, both at the organism and tissue/cell type level. We further show that eQTLs and especially dsQTLs in specific tissues tend to be significantly enriched among variants with high GenoNet scores, and how the GenoNet scores can help in the discovery of disease associated genes through an integrative analysis of lipid phenotypes using a Metabochip dataset on 12,281 individuals. In terms of cell type/tissue transferability of prediction models, we show using MPRA data in three cell lines that GenoNet trained on variants in one cell line has substantially reduced prediction accuracy on variants in a different cell line. As more systematic and comprehensive lists of experimentally validated variants become available across a large number of tissues and cell types over the next few years, GenoNet can be used to provide increasingly accurate functional predictions for variants genome-wide and across a variety of tissues and cell types.

Session 5: Basket Trials, Umbrella Trials, Platform Trials and Other Master Protocols

Overview of Basket and Umbrella Trials: Features, Challenges, and Examples

Lindsay Renfro
Mayo Clinic
renfro.lindsay@mayo.edu

Within the field of clinical cancer research, discovery of biomarkers and genetic mutations that are potentially predictive of treatment benefit are motivating a paradigm shift in how cancer clinical trials are conducted. In this talk, I will provide an overview of master protocols with special emphasis on basket and umbrella trials, which are increasingly popular design solutions for the study of novel targeted agents. For each, I will describe standardized terminology, discuss scientific and practical advantages and limitations, and provide a detailed real-world example.

Statistical Designs for Histology Agnostic Clinical Trials in Oncology

Richard Simon
R Simon Consulting
rmacesyimon@gmail.com


Use of Historical Control Data in Platform Trial
Satrajit Roychoudhury
Pfizer Inc.

Platform trials have the potential to drive tremendous efficiencies in the development of new therapeutics. These efficiencies can be operational or statistical in nature. Statistical efficiencies are derived from many sources including use of a shared control arm and adaptive randomization. In rare or difficult to enroll populations platform trials offer the opportunity to steer important new therapies into a standing clinical trials infrastructure potentially expediting the availability of badly needed new therapies. Efficiency can also be gained from statistical modeling to formally leverage both historical information and concurrent information in related populations. This framework allows the use of all relevant (historical and concurrent) data, for the inference of the parameter in the actual trial. In this talk, the various sources of efficiency will be described and real life examples will be provided to demonstrate their relative contribution to a more economical paradigm for clinical trials.

Basket Trial Designs Using Hierarchical Modeling and Dirichlet Process Priors

Mithat Gonen, Kristen Cunanan, Ronglai Shen and Alexia Iasonos
Memorial Sloan Kettering Cancer Center

Basket trials have quickly become one of the popular designs for developing targeted agents in precision oncology. They are originally favored for their logistical advantages of having a single protocol which allows treatment of patients with multiple disease types with the same agent. It has also become quickly clear that they can be statistically more efficient in identifying baskets where the treatment works. The gain efficiency comes from sharing information across baskets, which can be done in multiple ways. In this presentation we will focus on single-arm Phase II basket trials and present designs based on hierarchical models and Dirichlet process priors, comparing their operating characteristics with standard two-stage designs.

Session 6: Recent Work on Clinical Trial Design and Analysis

The current status and challenges of clinical efficiency assessment in Traditional Chinese Medicine
Baoyan Liu
Chinese Academy of Traditional Chinese Medicine Sciences

Treatment based on syndrome differentiation is the basic principle in traditional Chinese medicine and is a dynamic and individualize treatment. In current, the common clinical efficacy method is based on the clinical epidemiology methods and assesses the “simple paradigm”, which is not appropriate to the efficacy assessment to the treatment based on syndrome differentiation. We conducted series studies, including retrospective study, different TCM clinician cohort study, and randomization controlled trial to assess the clinical efficacy. In this study, we proposed a new opinion to assess the efficacy of TCM clinicians, not the prescriptions, according to the individual and complex intervention character of TCM. We also built a new TCM efficacy assessment methods, model and technological platform. Our study showed that the real world clinical research paradigm was a new method to assess the efficacy of TCM.

A Randomized Controlled Clinical Trial for Traditional Chinese Medicine Treatment of Primary Insomnia
Shiyian Yan
Institute of Clinical Basic Medicine, China Academy of Chinese Medical Science

Background Chinese medicine (CM) is popular for primary insomnia in China and beyond; however, its effects are uncertain. Methods We conducted a randomized, multicenter, triple-blind, parallel-group, placebo controlled trial involving 116 patients with primary insomnia. Three clinicians independently prescribed treatments for each patient based on syndromes differentiation. After one week washout period, patients were randomized to placebo or prescriptions by one of three clinicians (CM group) for a four-week treatment and a follow-up four weeks after. The primary outcome was change in total sleep time (TST) from baseline. Secondary endpoints included sleep onset latency (SOL), wake time after sleep onset (WASO), sleep efficiency, Pittsburgh Sleep Quality Index (PSQI) and CM symptoms. Adverse events (AEs) were monitored. Control group was stopped based on a preplanned interim analysis per protocol. Results The CM group had an average 0.62 more hours (95% Confidence Interval (CI): 0.30-0.94, P=0.0001) in TST increment and 34.14% (10.33%-57.96%, P=0.0066) more patients beyond 0.5 hour TST increment than that of the placebo group. PSQI was changed -3.28 (-3.82 to -2.74) in the CM group, a -2.01(-3.18 to -0.84, P=0.0011) difference from the placebo group. The CM symptoms score in the CM group decreased -1.99 (-3.28 to -0.69, P=0.0028) more than the placebo group. SOL, WASO, Beck Depression Inventory, and Self-rating Anxiety Scale score changes were not significantly different between groups. Interim and final analyses were consistent in every outcome, except SOL. The CM improvements in TST and PSQI were maintained at follow-up. Six AEs were reported, but all mild and unrelated to the treatment. Conclusions CM effectively and safely treats primary insomnia.

Inferring a null or non-null effect for binary endpoints in randomized blinded clinical trials
Li-Hsin Chien, Hua Jin, Jitendra Ganju and Ying Lu
Department of Biomedical Data Science and Center for Innovative Study Design, Stanford University
South China Normal University and Stanford University
Consultant, San Francisco, CA, USA

Zhou and Ganju [1] demonstrated that for continuous endpoints it is possible to reveal a null-non treatment effect in a blinded analysis when the randomization block size is known. In the current work, we develop a method to detect a null or a non-null effect for binary outcomes when the block sizes are 2 and 4. For a single binary endpoint, the power, as expected, is very low. However, because trials collect data on several binary endpoints, the power that at least one
endpoints detects the presence or absence of an effect is higher. We show how maximum likelihood estimates and confidence intervals of treatment effects can be constructed. Although the power is low, inferences regarding superiority, non-inferiority or equivalence are possible. Some simple steps are proposed to make inference difficult.


Session 7: Statistical inferences in some semi-parametric models

Longitudinal Mixed Effects Model with Fused Lasso Regularization

Judy Zhong
NYU Langone Health
judy.zhong@nyumc.org

The complex relationships between age and other risk factors produce highly variable natural histories from normal cognition to cognitive impairment. Many studies have shown the importance of modeling time-varying risk factors in predicting cognitive functioning. In this talk, we propose a mixed effect regression model with fused lasso regularization method to estimate the time-varying coefficients in longitudinal data. The constructed coefficient estimates are sparse and time-varying with adaptively selected change points. We will also introduce an efficient algorithm based on gradient descent for parameter estimation of this model especially in high dimensional settings. We use simulations to compare the proposed method with mixed effects regression models. We apply our proposed method to examine the longitudinal associations between clinical characteristics/self-reported chronic diseases with cognitive functioning among older Americans from the Health and Retirement Study.

Estimating Time-Varying Treatment Effect in Presence of Unmeasured Confounders

Haiqun Lin1 and Robert Rosenheck
1Yale University
haiqun.lin@yale.edu

Most currently available statistical methods assume that all the confounding variables between treatment and outcome are known, observed, or measured in order to make marginal inference in longitudinal data. However, in many circumstances, existence of unmeasured confounding is a rule rather than exception in observational studies or randomized studies with non-compliance. We develop a method of latent variable analysis of unmeasured confounding variables (LVAUC) for longitudinal data in this paper. The LVAUC approach aims at estimating effects attributable to a possibly time-varying treatment condition in the presence of time-varying (un)observed confounding variables. The estimator of treatment effect derived from the method is shown to be consistent and the performance of the method is tested with a simulations study. Our work extends marginal structural model proposed by Robins and colleagues by using weights that are estimated from joint modeling the relationship between the longitudinal outcome and the time-varying treatment. The method is also illustrated with a data set from a health services research study entitled Access to Community Care and Effective Services and Supports (ACCESS) program and evaluated by a simulation study.

Online clustering using Gaussian mixture models

Lei Yang

New York University
lei.yang@msn.com

Gaussian mixture models (GMM) are often used for data clustering. For large-scale data, the computational burden is heavy for conducting GMM clustering. Recently, online learning has been shown to be a scalable technique in areas of machine learning where it is computationally infeasible to train over the entire dataset. In this paper, we consider three online clustering approaches, (1) online K-means; (2) GMM via stochastic gradient descent (referred to as GMM-SGD); and (3) a hybrid approach that combines the strength of the other two approaches (referred to as GMM-hySGD). We demonstrate that online K-means is computationally fast, but it is not efficient; whereas GMM-SGD is efficient, but the speed is slowed down by applying the Riemannian manifold optimization. Instead, GMM-hySGD is efficient and computationally fast. We also incorporate the online selection of the number of clusters into the three approaches. We derive some asymptotic property on function values for GMM-hySGD because the corresponding objective function is non-convex. We examine the performance of the three approaches via simulation studies and a real data application.

Single-index model for inhomogeneous spatial point processes

Ji Meng Loh1 and Yixin Fang
1New Jersey Institute of Technology
loh@njit.edu

I will introduce a single-index model for the intensity of an inhomogeneous spatial point process, relating the intensity function to an unknown function \( \phi \) of a linear combination of a p-dimensional spatial covariate process. Such a model extends and generalizes the commonly used log-linear model. I will describe an estimating procedure for \( \phi \) and for the coefficient parameters \( \beta \). Consistency and asymptotic normality of estimates of \( \beta \) can be achieved under some regularity assumptions. I will show results from a simulation study showing the effectiveness of the procedure and from fitting the model to a dataset of fast food restaurant locations in New York City.

Session 8: Recent Advances in Dose Response Clinical Trials

A Cautionary Note When a Dose Ranging Study is Used for Proving the Concept

Qiqi Deng1, Kun Wang2, Xiaofei Bai3 and Naitee Ting1
1Boehringer Ingelheim Pharmaceuticals, Inc.
2Shenzhen Middle School
qiqi1015@hotmail.com

The objective of a Proof-of-Concept (PoC) clinical trial is to formulate a "Go/NoGo" decision based on study results. If such a decision cannot be made from outcomes of a PoC trial, then this creates a situation of inconclusiveness. Inconclusiveness could lead to many undesirable consequences. Recently, many project teams are combining the PoC with dose-ranging studies. When this is the case, there are likely more potential of causing inconclusiveness. This presentation points out some of these risks, and hopes to caution project team members to consider these risks while designing a combined PoC and dose-ranging clinical trial. When studying the problem of inconclusiveness, the concept of minimally statistically significant difference is extended from a two-sample PoC setting to a combined PoC and dose-ranging trial where multiple dose groups are involved.
Dynamic Development Paths for Expanding a Proof-of-Concept (PoC) Study to Explore Dose Range
Qi Qi Deng, Xiaofei Bai and Naitee Ting
Boehringer Ingelheim Pharmaceuticals, Inc.
polly.x.bai@gmail.com

In Phase II clinical development of a new drug, the two most important deliverables are proof of concept (PoC), and dose ranging. Traditionally a PoC study is designed as the first Phase II clinical trial. In this PoC, there are two treatment groups - a high dose of the study medication, against the placebo control. After the concept is proven, the next Phase II study is a dose ranging design with many test doses. This manuscript proposes a two-stage design with the first stage attempting to generate an early signal of efficacy. If successful, the second stage will adopt a "Go Fast" plan to expand the current study and add lower study doses of the test drug to explore the efficacy dose range. Otherwise, a "Go Slow" strategy is triggered, and the study will stop at a reduced sample size with high dose and placebo only.

Phase II dose spacing based on PK recommendations
Junxian Geng, Shihai Lu and Naitee Ting
Boehringer Ingelheim Pharmaceuticals, Inc.
junxian.geng@boehringer-ingelheim.com

In the process of new drug discovery and development, finding the right dose or the right range of doses is one of the most important objectives. In designing early Phase 2 dose response studies, one critical question is: "Which doses should be studied?" In this talk, we propose a dose allocation method guided by dose-exposure modeling. The relationship between the PK parameters such as AUC, Cmax and the dose levels are usually explored in Phase I studies. Simulations are performed to compare the proposed method with binary dose spacing method. Different measures are discussed while making the comparison. The proposed method for dose allocation can be applied to various types of dose response studies.

Session 9: Statistics and machine learning at the frontier of data-driven application and technology

Reinforcement learning for taxi driver dispatching
Zhiwei Qin
DiDi Research America
qin.zhiwei@didichuxing.com

In this talk, we first describe the challenging taxi driver dispatching problem on the DiDi ride-sharing platform. We will then lay out the evolution of solutions to this problem, from the existing combinatorial optimization-based solution to the more recent reinforcement learning methods. In particular, we will focus on a tabular temporal-difference (TD) learning method and the deep Q-networks (DQN) and how we construct the evaluation environment using real-world spatio-temporal trips data and train dispatching agents from such data. Due to the problem diversity across different cities, transfer learning is brought in to help increase the learning adaptability and efficiency. We empirically evaluating the performance of our dispatching algorithm and show the benefits of knowledge transfer in the spatial domain.

Extracting data from tables and charts in natural document formats
David Rosenberg
Bloomberg / New York University
david.davidsr@gmail.com

Financial analysis depends on accurate financial data, and these data are often distributed via PDF and other "natural document" formats. While these formats are optimized for easy human comprehension, automatically extracting the data can be quite challenging. We'll describe our work using a deep learning pipeline to extract data from tables and charts in PDF documents. We'll also show some of our latest research, inspired by image captioning models, for directly going from images of tables to a markup language (LaTeX) representation.

Data Science in e-commerce
Donghai Yan
University of Massachusetts Dartmouth
d@yan@umassd.edu

Data science has been playing increasingly important role in e-commerce. In this talk, I will introduce some data science applications that underlie the e-commerce, including personalization and item recommendation etc.

Session 10: Data-Driven Decision Making and Objective Statistical Modeling with Big Data Applications

Variable Selection Using Orthogonal Greedy Algorithms in Sparse High-Dimensional Time Series Models
Ching-Kang Ing
National Tsing Hua University
cking@stat.nthu.edu.tw

We investigate prediction capability of the orthogonal greedy algorithm (OGA) in high-dimensional time series models. We derive the prediction error rate of OGA in terms of the number of iterations and the number of candidate variables, under various sparsity conditions. We further introduce a high-dimensional model selection method, high-dimensional Akaike’s information criterion (HDAIC), to determine the number of OGA iterations. We show that when used together with HDAIC, OGA can achieve the desired error rate, which is sometimes better than the minimax-optimal rate, e.g., Raskutti et al. (2011) and Negahban et al. (2012), in the special case where observations are independent over time.

A New Information Criterion for Model Selection
Jie Ding1, Vahid Tarokh1 and Yuhong Yang2
1Duke University
djrlthu@gmail.com
2University of Minnesota

We introduce a new information criterion for model selection. It has the benefits of the two well-known model selection techniques, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). For a well-specified model class, BIC is typically consistent, and so is the new criterion. For a mis-specified model class, AIC is known to be asymptotically efficient in the sense that its predictive performance is asymptotically equivalent to the best offered by the candidate models; in this case, the new criterion behaves in a similar manner. While the optimality of AIC and BIC is susceptible to model specification, the proposed criterion can achieve the universal optimality in the sense that it can be automatically consistent in well-specified settings, and asymptotically efficient in mis-specified settings. In practice where the observed data is given without any prior information about the model specification, the proposed criterion can be more flexible and reliable compared with classical approaches. We also extend the criterion to high-dimensional regression settings where sample size is smaller than variable size.
Estimation and Optimization of Composite Outcomes

Daniel Luckett1, Eric Laber2 and Michael Kosorok1
1University of North Carolina at Chapel Hill
2North Carolina State University
luckett@live.unc.edu

There is tremendous interest in precision medicine as a means to improve patient outcomes by tailoring treatment to individual characteristics. An individualized treatment rule formalizes precision medicine as a map from patient information to a recommended treatment. A rule is defined to be optimal if it maximizes the mean of a scalar outcome in a population. However, clinicians often must balance multiple and possibly competing outcomes. One approach to precision medicine in this setting is to elicit a composite outcome which balances all competing outcomes; unfortunately, eliciting a composite outcome from patients is difficult without a high-quality instrument. We consider estimation of composite outcomes using observational data under the assumption that clinicians are approximately making decisions to maximize individual patient utility. Estimated composite outcomes are used to construct an estimator of a treatment rule that maximizes the mean of patient-specific composite outcomes. We prove that the proposed estimators are consistent and demonstrate their finite sample performance through simulation experiments and an application to a study of bipolar depression.

Efficient Online Bandit Multiclass Learning with sqrt(T) Regret

Chicheng Zhang
Microsoft Research
zcc1307@gmail.com

We present an efficient second-order algorithm with tilde O(sqrt(T)/η) regret for the bandit online multiclass problem. The regret bound holds simultaneously with respect to a family of loss functions parameterized by η, for a range of η restricted by the norm of the competitor. The family of loss functions ranges from hinge loss (η=0) to squared hinge loss (η=1). This provides a solution to the open problem of (J. Abernethy and A. Rakhlin. An efficient bandit algorithm for sqrt(T) regret in online multiclass prediction? In COLT. 2009). We test our algorithm experimentally, showing that it also performs favorably against earlier algorithms.

Session 11: Dynamic Modeling and Machine Learning for Big Data from Large Cohort Studies

Nonparametric Estimation of Risk Tracking Indices for Longitudinal Studies

Xin Tian and Colin Wu
National Heart, Lung and Blood Institute
xtianx@nhlbi.nih.gov

Tracking a subject’s risk factors or health status over time is an important objective in long-term epidemiology studies with repeated measurements. An important issue of time-trend tracking is to define appropriate statistical indices to quantitatively measure the tracking abilities of the targeted risk factors or health status over time. We present a number of local and global statistical tracking indices based on the rank-tracking probabilities, which are derived from the conditional distribution functions, and propose a class of kernel-based nonparametric estimation methods. We demonstrate the application of the tracking indices using the body mass index (BMI) and systolic blood pressure (SBP) data from the Coronary Artery Risk Development in Young Adults (CARDIA) study. Statistical properties of the estimation methods and bootstrap inference are investigated through a simulation study and an asymptotic development.

Cardiovascular Health Across the Lifespan

Colin Wa1 and Norrilla Allen2
1National Heart, Lung and Blood Institute
2Northwestern University
wuc@nhlbi.nih.gov

This study used the Lifetime Risk Pooling Project including African American and Caucasian participants of CARDIA, MESA, Framingham, ARIC, JHS and CHS. Individuals’ demographic data (age, birth year, sex, race, SES), traditional CV risk profile (smoking status, total cholesterol, SBP, DBP, glucose, lipid-lowering meds, antihypertensive meds, diabetes) and outcomes (incident CVD event and death) at each exam were included in the analysis. We proposed to impute the participant’s lifespan CV risk profile and outcome using the available exam records with the method of multi-level joint modeling multiple imputation. To validate our imputation, we removed the observed CARDIA data with exam age 18-30, MESA date with exam age 50-59 and CHS data with exam age 80-90. With the implement of jomo package in R, we imputed the CV risk profile, outcome at each of the observed age and then compared our imputed to the observed values. This synthetic cohort approach provides valid and unbiased estimates of CV risk factors across the lifespan. Future studies using this synthetic cohort can provide novel insights into the origins and accumulation of CV disease.

Local Box-Cox transformation on time-varying parametric models for smoothing estimation of conditional CDF with longitudinal data

Mohammed Chowdhury, Colin Wu and Reza Modarres
1Kennesaw State University
2National Heart, Lung and Blood Institute
3George Washington University
mchowd10@kennesaw.edu

Nonparametric estimation and inference of conditional distribution functions with longitudinal data have important applications in biomedical studies, such as epidemiological studies and longitudinal clinical trials. Estimation approaches without any structural assumptions may lead to inadequate and numerically unstable estimators in practice. We propose in this paper a nonparametric approach based on local Box-Cox transformation for estimating conditional distribution functions with a longitudinal sample. Our model assumes that the conditional distribution of the outcome variable at each given time point can be approximated by a parametric model after local Box-Cox transformation. Our estimation is based on a two-step smoothing method, in which we first obtain the raw estimators of the conditional distribution functions at a set of disjoint time points, and then compute the final estimators at any time by smoothing the raw estimators. Applications of our two-step estimation method have been demonstrated through a large epidemiological study of childhood growth and blood pressure. Finite sample properties of our procedures are investigated through a simulation study. Application and simulation results show that smoothing estimation from time-varying parametric models outperforms the existing kernel smoothing estimator by producing narrower pointwise bootstrap confidence band and smaller root mean squared error.

Keynote lecture II

Keynote lecture II: The Changing Landscape of Drug Development: To Adapt or Not to Adapt

Dionne Price
Office of Biostatistics, Office of Translational Sciences Center for Drug Evaluation and Research (CDER), FDA

Statisticians may need to adapt to meet the demands of the changing landscape of drug development. This need to adapt has been motivated by numerous factors including the new era of big data as well challenging problems in various therapeutic areas. The Food and Drug Administration is poised to meet the demands of the changing landscape as evident by various commitments outlined in recent legislation such as the 21st Century Cures Act and the Prescription Drug User Fee Act (PDUFA) reauthorization for fiscal years 2018-2022, known as PDUFA VI. These commitments encompass a broad range of topics including real world evidence, complex innovative designs, patient-focused drug development, and model-informed drug development. To achieve the common goal of advancing drug development to meet the evolving needs of patients, statistical input, collaboration, and leadership will be germane. This talk will provide an overview of the commitments and discuss how the Office of Biostatistics is adapting to propel drug development forward in this new era.

**Session 12: Deep learning and its mechanism**

**A Unified View of Deep Generative Models**

*Eric Xing*

Carnegie Mellon University

We present a unifying theoretical framework that connects GANs, VAEs, with the classical wake-sleep algorithm, and the variational inference method for latent variable models, and show that these methods, as well as many other variational models invented in recent years, can be easily formulated as instances or approximations of a loss-augmented posterior inference problem of latent variable graphical models. I show how such a unified view can help modelers create new DGM variants for different applications, and derive standard learning/inference algorithms.

**A Geometric View to Optimal Transportation and Generative Model**

*Xianfeng Gu*

Stony Brook University

Generative Adversarial Net (GAN) is a powerful machine learning model, and becomes extremely successful recently. The generator and the discriminator in a GAN model competes each other and reaches the Nash equilibrium. GAN can generate samples automatically, therefore reduce the requirements for large amount of training data. It can also model distributions from data samples. In spite of its popularity, GAN model lacks theoretic foundation. In this talk, we give a geometric interpretation to optimal mass transportation theory, and applied for the GAN model. We try to answer the following fundamental questions:

1. Does a GAN model learn a function, a mapping or a probability distribution? Is the solution unique or infinite? What is the dimension of the solution space? What is the structure of the solution space?
2. Does a GAN model really learn or just memorize?

**Analysis on gradient descent dynamics in Deep Neural Network**

*Yuandong Tian*

Facebook AI Research

Recently, Deep Learning has made substantial progress on many disciplines that are traditional hard, such as computer vision, speech recognition and natural language processing. However, as a highly nonlinear formulation with high-dimensional non-convex optimization, its underlying mechanism is still unknown. In this talk we present some of our recent progress to reveal the dynamics and convergence behavior of gradient descent that is extensively used in training deep nonlinear networks.

Besides, we will also present our practical works on deep reinforcement learning platform and its applications to various games. This will be presented in Section #93.

**Session 13: Making sense of data for business: the role of statistics**

**Multilevel Log-Gaussian Cox Process Modelling for Structured Temporal Point Processes with Applications in Stock Market Trading**

*Yongtao Guan*

University of Miami

We propose a novel multilevel log-Gaussian Cox process modeling procedure for temporal point processes. The proposed procedure can be used to model repeatedly observed temporal point patterns which have become increasingly available due to technological advancement. A completely nonparametric approach is developed to consistently estimate the covariance kernels of the latent component processes at all levels. We further extend our procedure to the bivariate point process case, where potential associations between the processes can be assessed. Asymptotic properties of the proposed estimators are investigated and the effectiveness of our procedures is illustrated through a simulation study and an application to a stock trading dataset.

**Learning heterogeneity in causal inference using sufficient dimension reduction**

*Wenbo Wu*

University of Oregon

Often the research interest in causal inference is on the regression causal effect, which is the mean difference in the potential outcomes conditional on the covariates. In this paper, we use sufficient dimension reduction to estimate a lower dimensional linear combination of the covariates that can be used in three ways: to conduct variable selection for the regression causal effect, to improve the estimation accuracy of the regression causal effect, and to detect the heterogeneity of the regression causal effect. Compared with the literature, our approaches adopt a weaker sufficient dimension reduction assumption, and do not rely on parametric modeling of the regression causal effect or any modeling of the individual outcome regres-
Threshold Factor Models for High-Dimensional Time Series
Xialu Liu
San Diego State University
xialu.liu@sdsu.edu

We consider a threshold factor model for high-dimensional time series in which the dynamics of the time series is assumed to switch between different regimes according to the value of a threshold variable. This is an extension of threshold modeling to a high-dimensional time series setting under a factor structure. Specifically, within each threshold regime, the time series is assumed to follow a factor model. The factor loading matrices are different in different regimes. The model can also be viewed as an extension of the traditional factor models for time series. It provides flexibility in dealing with situations that the underlying states may be changing over time, as often observed in economic time series and other applications. We develop the procedures for the estimation of the loading spaces, the number of factors and the threshold value, as well as the identification of the threshold variable. We investigate their theoretical properties, and find that even when the number of factors is overspecified, our estimators are still consistent and converge as fast as these when the number of factors is correctly specified. Simulated and real data examples are presented to illustrate the performance of the proposed method.

Structural Break Detection in Financial Durations
Yaohua Zhang
Vertex Pharmaceuticals
yaohua.zhang@vrtx.com

High-frequency financial data are more readily available and provide a deeper understanding of market infrastructure, market dynamics and structural instability. The class of autoregressive conditional duration (ACD) models are useful for statistical analysis of intra-event durations in asset prices. Often, time series of durations exhibit structural breaks that may be due to change in level, or change in model order, or change in parameter values. In this article, we study the structural break detection problem in univariate time series of durations using penalized estimating functions (PEFs). We propose a retrospective algorithm based on the entire data that allows us to simultaneously detect the number of structural breaks, locations of the structural breaks as well as the model order. We present a simulation based comparison between our algorithm and a Group LASSO method that has been discussed in the literature for piecewise autoregressive (AR) models. Our method based on the PEF is attractive because it provides insights for understanding the stochastic dynamics of high-frequency financial data.

Session 14: Challenges and Innovations in Modeling Large-Scale Imaging Data

A time-varying AR coefficient model of functional near-infrared spectroscopy data
Marco Benedetti and Timothy Johnson
University of Michigan
benedetti@umich.edu

Functional near-infrared spectroscopy (fNIRS) is a relatively new neuroimaging technique. It is a low cost, portable, and non-invasive method to monitor brain activity. Similar to fMRI, it measures changes in the level of blood oxygen in the brain. Its time resolution is much finer than fMRI, however its spatial resolution is much coarser—similar to EEG or MEG. fNIRS is finding widespread use on young children whom cannot remain still in the MRI magnet and it can be used in situations where fMRI is contraindicated—such as with patients whom have cochlear implants. In this talk, I propose a fully Bayesian time-varying autoregressive model to analyze fNIRS data. The hemodynamic response function is modeled with the canonical HRF and the low frequency drift with a variable B-spline model (both locations and number of knots are allowed to vary). Both the model error and the auto-regressive process vary with time. Via a simulation studies, I show that this model naturally handles motion artifacts and gives good statistical properties. The model is then applied to a fNIRS study.

Can neuroimaging diagnose neurodegenerative diseases?  
Lei Wang1, Karteek Popuri, Rakesh Balachandar, Kathryn Alpert, Donghuan Lu, Mahadev Bhalla, Ian Mackenzie, Robin Hsiung, Veronika Hanco, Kristen Warren, Konstantinos Arfanakis, Julie Schneider, David Bennett and Mirza Faisal Beg
1 Feinberg School of Medicine, Northwestern University
leiwang1@northwestern.edu

Despite advances in the development of biomarkers for Alzheimer’s disease (AD), accurate ante-mortem diagnosis remains challenging. Incorporating biomarker evidence for AD pathology can increase diagnostic certainty. Here we report our efforts through two studies: 1) utilizing machine learning on multimodal ante-mortem big data to develop FDG-PET biomarkers; 2) utilizing postmortem neuropathology data to develop ante-mortem hippocampal predictors of AD neuropathology burden. In the first study, we used a multi-scale ensemble machine learning to train and validate a classification model on fluorodeoxyglucose positron emission tomography (FDG-PET) imaging data. The model was trained on 3D topographic brain glucose metabolism patterns from normal controls (NC) and individuals with dementia of Alzheimer’s type (DAT), and validated on individuals with mild cognitive impairment (MCI) who converted to DAT. The experiments were performed on a large number of FDG-PET images (N=2,984) obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database. Conversion from mild cognitive impairment (MCI) to DAT was predicted with an AUC of 0.81, 0.80, 0.77 for the 2, 3, 5 years to conversion windows respectively. We also computed a FDG-PET DAT score (FPDS, scalar between 0 and 1) to indicate the probability of the brain metabolism pattern for DAT. These results indicate that multi-scale supervised ensemble learning on FDG-PET data can be used to probabilistically predict clinical DAT status. In the second study, we evaluated the potential of using ante-mortem hippocampal surface shape measures to predict neuropathologies. In 42 participants from longitudinal cohort studies, surfaces were generated for the hippocampus in their ante-mortem T1-weighted MRI, and postmortem neuropathology burdens (counts) for amyloid-beta, PHF-tau, and TDP-43 were evaluated in the whole brain and in the hippocampus. We mapped the relationships between measures of hippocampal shape and neuropathology onto the hippocampal surface using multiple linear regression models. Inward hippocampal deformity in the CA1 and subiculum regions was predicted by higher PHF-tau burden, but no significant patterns of inward surface deformity were associated with amyloid-beta or TDP-43 after including covariates. These results indicate that hippocampal CA1 atrophy may represent a biomarker for postmortem AD neuropathology.

Challenges in the analysis of MRI and PET data in studies of Alzheimer’s disease and Down syndrome
Dana Tudorascu1, Davneet Minhas1, Zheming Yu1, Ben Handen1

Abstracts
Making them difficult to interpret. In this talk, I will propose a large percentage of identified variants reside in non-coding regions, variants with weak effect sizes. Second, and more importantly, a increasing sample sizes, these studies are still underpowered for...
are often involved in condition-specific activities. Interestingly, we observe that the genes strongly associated with survival time in the TCGA dataset are less likely to be network hubs, suggesting that genes associated with cancer progression are likely to govern specific functions, rather than regulating a large number of biological processes. Additionally, we observed that the tumor-specific hub genes tend to have few shared edges with normal tissue, revealing tumor-specific regulatory mechanism.

**Novel method for inferring genes that escape X chromosome inactivation from RNA seq data**

Dajiang Liu
Pennsylvania State University
[dxl145@psu.edu](mailto:dxl145@psu.edu)

X chromosome inactivation (XCI) mechanism randomly silences one female X (Xi) and only expresses the alleles on the active X (Xa). 10% of the X-linked genes escape XCI and express both copies of alleles. Xa/Xi assignment varies between cells (mosaicism), which makes it challenging to identify XCI escape genes from bulk RNA-seq data.

We propose a novel binomial mixture model for identifying XCI escape gene. The method proceeds by estimating XCI skewing (i.e. the fraction of cells where one particular haplotype is Xa) from a training set of putative inactivated genes. The genes with more balanced allelic expression than XCI skewing will be deemed as candidates of XCI escape genes. The model accounts for the read depths over-dispersion, while simultaneously accommodating for possible sequencing error or imperfect training set (e.g. some putative inactivated genes may actually escape XCI). The number of mixture components is automatically determined via a backward elimination procedure, which ensures model parsimony and maximizes power.

We experimentally established the validity and power of proposed method using single cell clones and male-female differential gene expression (DGE). In the single cell clone experiment, cells from a single individual with maternal or paternal Xa were isolated and separately cloned. Each clone thus has identical Xa/Xi assignment. Any genes that are mono-allelically/bi-allelically expressed in single cell clones will be identified as inactivated/escape genes. We also created mixtures of single cell clones, with skewing ranging from 50/50 to 90/10 to mimic a randomly X-inactivated sample. Type-I error and power were evaluated by comparing inference results from mixture samples to single cell clones. The type I errors across all scenarios were well controlled. The power increases with XCI skewing and is well above 80% for cell mixtures with a skewing greater than 20/80.

We further demonstrated the utility of this method with DGE in the GEUVADIS dataset. Our method correctly predicted all 11 genes in PAR1 region as escape gene. Consistent with existing literature, the genes in non-PAR region predicted as escape tend to be up-regulated in females. On the contrary, genes in the PAR1 region are down-regulated in females due to dosage compensation. The method is implemented as a publicly available R package. We envision that it will be extremely useful for XCI escape analysis using broadly available bulk RNA-seq data.

**Simulation Studies of Two Dose Escalation Methods for Oncology Drug Combination Therapies**

Jing Hu¹, Ron Yu² and Kc Huang²
¹Genentech
²Gilead Sciences
[jasmine.jing.hu@gmail.com](mailto:jasmine.jing.hu@gmail.com)

To address the increasing need for more efficient designs, especially for drug combination studies, we evaluated two promising model-based dose escalation methods: Bayesian logistic regression model (BLRM) method and product of independent beta probabilities escalation (PIPE) method in the literature. Eight different simulation scenarios were designed to compare the performance of these two methods. The simulation results and the findings will be discussed.

**Benefits of Borrowing Information in Oncology Basket Trial**

Fang Liu and Cong Chen
Merck & Co., Inc.
[fang.liull@merck.com](mailto:fang.liull@merck.com)

As cancer treatments have entered in a new era of immunoncology, strategies have been developed over the last several years to help expedite the drug development process and ensure that patients receive the most appropriate therapies. One of these strate-
Evaluating accuracy of diagnostic tests without conditional information in a spinal cord injury data set.

Comparing time to functional progression between groups defined initially introducing that our proposed test remains unbiased under cluster-based stratifying. In this talk, we introduce a log rank test for clustered data that adjusts for the population size and informative within-cluster group size. In this talk, we investigate via simulation the potential benefits of several information borrowing approaches comparing to a simple independent evaluation (by tumor type) approach.

Session 17: Recent Advances in Nonparametric Methods

Rank-Based Tests for Clustered Data with R package clusrank

Yujing Jiang1, Jun Yan2, Xin He3, Mei-Ling Ting Lee3 and Bernard Rosner4
1Colorado State University
2University of Connecticut
3University of Maryland, College Park
4Harvard University

Wilcoxon Rank-based tests are distribution-free alternatives to the popular two-sample and paired t-tests. For independent data, they are available in several R packages such as stats and coin. For clustered data, we present a package, clusrank, where the latest developments are implemented and wrapped under a unified user-friendly interface. With different methods dispatched based on the inputs, this package offers great flexibility in rank-based tests for various clustered data. Exact tests based on permutations are also provided for some methods. Usages of clusrank are illustrated with simulated data.

A log rank test for clustered data with informative within-cluster group size

Somnath Datta1, Mary Gregg2 and Doug Lorenz2
1University of Florida
2University of Louisville

The log rank test is a popular nonparametric test for comparing survival distributions among groups. When data are organized in clusters of potentially correlated observations, adjustments can be made to account for within-cluster dependencies among observations, e.g., tests derived from frailty models. Tests for clustered data can be further biased when the number of observations within each cluster and the distribution of groups within cluster are correlated with survival times, phenomena known as informative cluster size and informative within-cluster group size. In this talk, we introduce a log rank test for clustered data that adjusts for the potentially biasing effect of informative cluster size and within-cluster group size. We provide the results of a simulation study demonstrating that our proposed test remains unbiased under cluster-based informativeness, while other candidate tests not accounting for the clustering structure do not properly maintain size. Further, our test exhibits power advantages under scenarios in which traditional tests are appropriate. We demonstrate an application of our test by comparing time to functional progression between groups defined initial functional status in a spinal cord injury data set.

Evaluating accuracy of diagnostic tests without conditional independence assumption

Larry Tang

George Mason University
ltang@gmu.edu

Evaluating the accuracy (ie, estimating the sensitivity and specificity) of new diagnostic tests without the presence of a gold standard is of practical meaning and has been the subject of intensive study for several decades. Existing methods use 2 or more diagnostic tests under several basic assumptions and then estimate the accuracy parameters via the maximum likelihood estimation. One of the basic assumptions is the conditional independence of the tests given the disease status. This assumption is impractical in many real applications in veterinary research. Several methods have been proposed with various dependence models to relax this assumption. However, these methods impose subjective dependence structures, which may not be practical and may introduce additional nuisance parameters. In this article, we propose a simple method for addressing this problem without the conditional independence assumption, using an empirical conditioning approach. The proposed method reduces to the popular Hui-Walter model in the case of conditional independence. Also, our likelihood function is of order-2 polynomial in parameters, while that of Hui-Walter is of order-3. The reduced model complexity increases the stability in estimation. Simulation studies are conducted to evaluate the performance of the proposed method, which shows overall smaller biases in estimation and is more stable than the existing method, especially when tests are conditionally dependent. Two real data examples are used to illustrate the proposed method.

The noise barrier and the large signal bias of the Lasso and other convex estimators

Pierre Bellec
Rutgers University
pcb71@stat.rutgers.edu

Convex estimators such as the Lasso, the matrix Lasso and the group Lasso have been studied extensively in the last two decades, demonstrating great success in both theory and practice. Two quantities are introduced, the noise barrier and the large scale bias, that provides insights on the performance of these convex regularized estimators. It is now well understood that the Lasso achieves fast prediction rates, provided that the correlations of the design satisfy some Restricted Eigenvalue or Compatibility condition, and provided that the tuning parameter is large enough. Using the two quantities introduced in the paper, we show that the compatibility condition on the design matrix is actually unavoidable to achieve fast prediction rates with the Lasso. The Lasso must incur a loss due to the correlations of the design matrix, measured in terms of the compatibility constant. This results holds for any design matrix, any active subset of covariates, and any tuning parameter. It is now well known that the Lasso enjoys a dimension reduction property: the prediction error is of order $\lambda \sqrt{k}$ where $k$ is the sparsity; even if the ambient dimension $p$ is much larger than $k$. Such results require that the tuning parameters is greater than some universal threshold. We characterize sharp phase transitions for the tuning parameter of the Lasso around a critical threshold dependent on $k$. If $\lambda$ is equal or larger than this critical threshold, the Lasso is minimax over $k$-sparse target vectors. If $\lambda$ is equal or smaller than critical threshold, the Lasso incurs a loss of order $\sigma \sqrt{k}$ which corresponds to a model of size $k$—even if the target vector has fewer than $k$ nonzero coefficients. Remarkably, the lower bounds obtained in the paper also apply to random, data-driven tuning parameters. The results extend to convex penalties beyond the Lasso.
Session 18: Advancing Complex Innovative Clinical Trial Design

Advancing Complex Innovative Clinical Trial Design
Freda Cooner1 and Fanni Natanegara
1Sanofi
Freda.cooner@sanofi.com
The cost for bringing a potential treatment from discovery to market has nearly doubled in the last 10 years and has now reached an astronomical amount of $2.5 billion. On average, it takes at least ten years for a new medicine to complete the journey from initial discovery to the marketplace, with clinical trials alone taking six to seven years on average. There is an urgent need to make drug development less costly and less time consuming. Innovative trial designs/analyses such as complex adaptive and Bayesian approaches are essential to meet this need. Recent regulations such as the 21st Century Cures Act and PDUFA VI have the potential to dramatically change the landscape for use and acceptance of complex innovative trial designs in drug and biological product development to help inform regulatory decision making. These policies specifically call out the broader application of Bayesian statistics and adaptive trial designs including issuing guidance documents, organizing public workshops, and conducting pilot programs to facilitate technical discussions between regulators and sponsors related to modeling and simulations while an innovative design is being planned. It is important for statisticians across industry, regulatory, and academia to understand the impact of these legislation, collaborate to ensure open communication and appropriate as well as consistent application of these innovative methods. Four invited panel members from industry and regulatory will be speaking at this session. We expect that this session will have a wide appeal due to the recency of these legislation and urgent need to accelerate the pace of treatments and cures.

Session 19: Real-World Evidence: Moving from Data to Knowledge and Decision Making

Transforming Real World Data into Scientific Evidence for the Regulatory Decision-Making
Vandana Mukhi
Center for Devices and Radiological Health, FDA
vandana.mukhi@fda.hhs.gov
Under certain circumstances real world data (RWD), collected in form of registries, can be used to support regulatory decisions for medical devices. When considering RWD in a regulatory submission, it is important that a plan is created prior to accessing, retrieving and analyzing RWD. In this presentation we provide examples of where registry data has been used. Statistical approaches that can be used to make valid statistical inference will also be discussed.

Treatment decision using causal survival forests
Yifan Cui1, Michael Kosorok1, Stefan Wager and Ruqing Zhu
1University of North Carolina at Chapel Hill
yifanc095@gmail.com
Survival forests with observational data may be quite biased in finite samples. Part of the reason for this problem is that, when we grow a survival forest, we need the recursive partitioning to simultaneously do many different things: express heterogeneity in our quantity of interest, mitigate confounding, and reduce variance. We propose an alternative approach to constructing survival forests that seeks to overcome these issues. We start with classical results on semiparametric efficiency in survival analysis and then use the formalism of generalized random forests (Athey et al 2017) to build a forest. The performance of the proposed approach is demonstrated via simulation studies.

Diagnosis of diabetic retinopathy using medical images and deep learning methods
Xuanyao He and Haoda Fu
Eli Lilly and Company
he_xuanyao@illy.com
Diabetic Retinopathy (DR) is a disease of the retina which affects patients with diabetes mellitus and is the main cause of blindness. This disease, in which the retinal blood vessels swell, has increasing prevalence among working-age population. The most effective treatment is early detection through regular/manual screenings. Automatic screening methods of DR using images with high accuracy have the potential to assist physicians in evaluating patients earlier, thereby potentially enabling patients to seek timely help from specialists. These methods emphasize on determination of retinal images using appropriate image processing and data mining techniques. In this presentation, we apply deep learning methods to classify a given set of Kaggle images into 5 distinct classes. The classification is carried out through transfer training model and neural networks. Examples will be provided to demonstrate the opportunity and ability of machine learning techniques to help solve this important medical problem.

Session 20: The role of real world evidence in the development of medical treatment

Use of Real World Evidence for premarketing evaluation of medical devices
Nelson T. Lu
FDA
Nelson.Lu@fda.hhs.gov
Various sources of real world data could be leveraged in the clinical studies in the regulatory settings. Such data may reflect real world clinical practice and could potentially be used to reduce the cost of the traditional clinical trials. However, challenges arise regarding how to transform the real world data into valid scientific evidence that can be used in the regulatory decision making for premarketing approval of medical device products. This talk will present some examples and discuss some considerations and challenges from the statistical and regulatory perspectives.

Applying propensity score adjustment in synthesizing real-world data in regulatory decision making
Chenguang Wang
Johns Hopkins University
cwang68@jhmi.edu
The amount of real-world data (RWD) collected from sources in real-world settings is increasing ultrarapidly. With the wide availability of various RWD, there has been a great level of interest in using real-world data in regulatory decision makings. In this paper, we explore the advantages and possible issues of using propensity score adjustment for incorporating RWD in the design and analysis of single group pre-market clinical trials. Numerical studies are conducted to evaluate the performance of different methods in different settings. In addition, a hypothetical single group medical device study is provided to illustrate the relevant methods in the context of regulatory review.
Weighting Strategies for Comparative Effectiveness Research using Observational Data
Julia Wang, Xiwu Lin and Kwan Lee
Janssen PRD
jwang@lts.jnj.com

In addition to stratification, matching, and regression adjustment, weighting methods are a popular strategy to remove covariates imbalance and confounding in observational studies. Under certain conditions, most notably unmeasured confounding, inverse probability weighting (IPW) based on estimated propensity scores using the Horvitz-Thompson rationale from survey sampling creates a pseudo superpopulation within which causal treatment effects can be estimated without bias. Implementation of IPW in practice often runs into difficulties with very large weights for some observations, causing unstable effect estimation and large variability. Recently, many new weights have been proposed, aiming to ameliorate the issues associated with IPW. These include Covariate-Balance Propensity Score (CBPS) based weight, overlap weight, matching weight, and trapezoidal weight. In this presentation, we will illustrate the strengths and weaknesses of various weighting strategies using simulation, and provide an understanding of the most suitable circumstance under which each weight should be used.

Session 21: Statistical methods and applications in mobile health

Partially Observed Dynamic Models for Tracking Therapeutic and Engagement Outcomes
Zhenke Wu1, Walter Dempsey2 and Susan Murphy3
1Department of Biostatistics, University of Michigan
2Department of Statistics, Harvard University
3Department of Statistics and of Computer Science, Harvard University
zhenkewu@umich.edu

In mobile health, a critical composite, dynamic, hypothetical latent construct is engagement with the mobile application. A joint model for the dynamic health behavior and engagement outcomes is required to inform their dynamic interrelationship. In this paper, we introduce a hierarchical, partially observable Gaussian process model that depends on dynamic exogenous variables and multiple noisy measures of the latent construct. The model accounts for interventions that can lead to increased step counts and decreased engagement as well as alter their correlation. We estimate the models in a Bayesian framework via Markov chain Monte Carlo algorithms which by design produce samples that approximate the posterior distribution of the unknowns given the observables. We illustrate the use of the model through analyses of data from HeartSteps, a mobile health intervention study aimed at increasing physical activity, and provide an understanding of the most suitable circumstance under which each weight should be used.

Compositional and functional data analysis methods for accelerometer data
Chongzhi Di and Yifan Zhu
Fred Hutchinson Cancer Research Center
cdi@fredhutch.org
yzhu2@fredhutch.org

Accelerometers are widely used to objectively measure physical activity in epidemiological studies. In these studies, it is often of interest to investigate whether activity of varying intensity levels, e.g., sedentary behavior (SB), light physical activity (LPA) and moderate and vigorous physical activity (MVPA), are independently related to health outcomes. However, these variables often have high collinearity. Several models have been used in the literature, but they yield conflicting conclusions. We propose compositional and functional data analysis approaches to solve the puzzle. The proposed approaches are flexible and robust, unless existing models that rely on restrictive assumptions. The methods are demonstrated by an application to the Women’s Health Initiative to investigate the association between physical activity and cardiometabolic biomarkers in 6,500 older women.

Clustering functional data with application to electronic medication adherence monitoring in HIV prevention trials
Yifan Zhu, Chongzhi Di and Ying Qing Chen
Fred Hutchinson Cancer Research Center
y2zhu@fredhutch.org

Maintaining high medication adherence is essential for achieving desired efficacy in clinical trials, especially prevention trials. However, adherence is traditionally measured by self-reports that are subject to reporting biases and measurement error. Recently, electronic medication dispenser devices have been adopted in several HIV pre-exposure prophylaxis prevention studies. These devices are capable of collecting objective, frequent, and timely drug adherence data. The device opening signals generated by such devices are often represented as regularly or irregularly spaced discrete functional data, which are challenging for statistical analysis. In this paper we focus on clustering the adherence monitoring data from such devices. We first pre-process the raw discrete functional data into smoothed functional data. Parametric mixture models with change-points, as well as several non-parametric functional clustering approaches are adapted and applied to the smoothed adherence data.

Simulation studies were conducted to evaluate finite sample performances, on the choices of tuning parameters in the pre-processing step as well as the relative performance of different clustering algorithms. We applied these methods to the HIV Prevention Trials Network (HPTN) 069 study for identifying subgroups with distinct adherence behavior over the study period.

Session 22: Recent Developments in Risk Prediction Models

Dementia risk prediction in preclinical stage
Zheyu Wang
Johns Hopkins University
wangzy@jhu.edu

The initiation of the Alzheimer disease (AD) pathogenic process is typically unobserved and has been thought to precede the first symptoms by 10 years or more. This impose a major challenge in investigating biomarkers for early AD detection, because 1) using clinical diagnosis as the reference point can be in error, especially in the early course of the disease; and 2) most AD studies do not have autopsy data to confirm diagnoses. Until technology advance allows for brain examination with “autopsy level” clarity, an appropriate statistical method that directly address the unobservable nature of preclinical AD progression is necessary for any rigorous AD biomarker evaluation and for efficient analyzing AD study data where only clinical data are available and neuropathology data are not yet available. This lack of gold standard issue is particularly prevalent in biomarker evaluation: the use of biomarker is often to facilitate early detection, to reduce cost, and/or to improve diagnostic accuracy, which means that current diagnostic standard may not be able to provide conclusive or accurate evidence in time, and may be too costly to conduct for all at risk individuals. In this talk,
we will discuss two latent variable models for this purpose. Most of the attention will focus on a model that considers categorical latent disease status and intends to assess biomarker’s utility in help making medical decision. We will also outline a model that considers continuous latent disease progression and intends to adopt biomarker information to help predict subjects’ AD progression trajectory. Applications to Alzheimer’s disease data will be discussed.

Risk Prediction for Heterogeneous Populations with Application to Hospital Admission Prediction

*Jared Huling*, Menggang Yu, Maxuan Liang and Maureen Smith

1Ohio State University

huling.78@osu.edu

There is an increasing need to model risk for large health systems that provide services to diverse and complex patients. Often, heterogeneity across health system populations can be explained by a set of factors such as chronic conditions. For example, patients with congestive heart failure often have fundamentally different needs and risks from patients with diabetes and furthermore from patients with both congestive heart failure and diabetes. However, taking such heterogeneity into account exacerbates estimation issues that arise from the high dimensionality of the data. We exploit the underlying structure of the different subpopulations by imposing structural constraints on the importance of variables. The structural assumptions are scientifically plausible and aid greatly in borrowing strength across the subpopulations. We prove an oracle property for our estimation method and show that it is robust to significant violations of our key structural assumption. We demonstrate the impressive performance of our method in extensive numerical studies and on an application in hospital admission prediction for the Medicare population of a large health system.

A Joint Model for Partial Peer Network Data with Informative Cluster Size

*Damping Liu*, Patrick Coyle, Denise Haynie, Bruce Simons-Morton and Joe Bible

1National Cancer Institute

2Temple University

3Eunice Kennedy Shriver National Institute of Child Health and Human Development

4Clemson University

danping.liu@nih.gov

In risk prediction, it is common to collect information of predictors that have a clustered or longitudinal structure. For example, NEXT Generation Health Study is interested in predicting alcohol use of adolescents, where participants together with their nominated close friends answer a survey questionnaire about their behaviors, providing a unique opportunity to understand peer influence in risky behaviors among the adolescents. A unique feature of the data is that, the number of peers each participant had was random; and conceivably, the number of nominated peers might be associated with the drinking outcome, as well as the strength of peer influence. Meanwhile, there were a number of participants who did not nominate any peers, introducing another potential complication from a biased sample. We develop a novel joint model to account these unique data features. The joint model has three components, a model for the participant’s outcome, a model for the peer outcome, and a model for the informative size of the peer network. A random effect term is shared in these model components to introduce a dependence among them. We discuss the advantage of this new method with several simple alternatives that ignore the informative network size, and compare their performance in a series of simulation studies. The risk predictions for alcohol use was compared across methods in the NEXT data example.

Session 23: classification for high dimensional complex data

Sparse quadratic classification rules via linear dimension reduction

*Irina Gaynanova and Tianying Wang*

Texas A & M University

irinag@stat.tamu.edu

We consider the problem of high-dimensional classification between the two groups with unequal covariance matrices. Rather than estimating the full quadratic discriminant rule, we propose to perform simultaneous variable selection and linear dimension reduction on original data, with the subsequent application of quadratic discriminant analysis on the reduced space. In contrast to quadratic discriminant analysis, the proposed framework doesn’t require estimation of precision matrices and scales linearly with the number of measurements, making it especially attractive for the use on high-dimensional datasets. We support the methodology with theoretical guarantees on variable selection consistency, and empirical comparison with competing approaches. We apply the method to gene expression data of breast cancer patients, and confirm the crucial importance of ESR1 gene in differentiating estrogen receptor status.

Dynamic linear discriminant analysis in high dimensional space

Binyan Jiang

The Hong Kong Polytechnic University

by.jiang@polyu.edu.hk

High-dimensional data that evolve dynamically feature predominantly in the modern data era. As a partial response to this, recent years have seen increasing emphasis to address the dimensionality challenge. However, the non-static nature of these datasets is largely ignored. This paper addresses both challenges by proposing a novel yet simple dynamic linear programming discriminant (DLPD) rule for binary classification. Different from the usual static linear discriminant analysis, the new method is able to capture the changing distributions of the underlying populations by modeling their means and covariances as smooth functions of covariates of interest. Under an approximate sparse condition, we show that the conditional misclassification rate of the DLPD rule converges to the Bayes risk in probability uniformly over the range of the variables used for modeling the dynamics, when the dimensionality is allowed to grow exponentially with the sample size. The minimax lower bound of the estimation of the Bayes risk is also established, implying that the misclassification rate of our proposed rule is minimax-rate optimal. The promising performance of the DLPD rule is illustrated via extensive simulation studies and the analysis of a breast cancer dataset.

Quadratic discriminant analysis for high-dimensional data

*Yingli Qin, Yilei Wu and Mu Zhu*

1University of Waterloo

yingli.qin@uwwaterloo.ca

High-dimensional classification is an important and challenging statistical problem. We develop a set of quadratic discriminant rules by simplifying the structure of the covariance matrices instead of imposing sparsity assumptions - either on the covariance matrices themselves (or their inverses), or on the standardized between-class distance. Under moderate conditions on the population covariance
matrices, our specialized quadratic discriminant rules enjoy good asymptotic properties. Computationally, they are easy to implement and do not require large-scale mathematical programming. Numerically, they perform well in finite dimensions and with finite sample sizes. We also present real-data analyses of several classic micro-array data sets.

Sparse Tensor Additive Learning for Click-Through-Rate Prediction
Will Wei Sun
University of Miami
wsun@bus.miami.edu
In online advertising, it is of interest to predict the user’s click-through-rate (CTR) given the ad tensor data. Preliminary analysis shows severe nonlinearity between the overall CTR and the input ad impression tensor, so that the linear tensor model is no longer applicable. This motivates us to consider an additive model for supervised tensor learning. Estimation accuracy and computational costs are evaluated and compared with existing linear and nonlinear tensor learning methods. An interesting statistical-and-computational tradeoff will be characterized through the error bound of the output estimate at each iteration.

Bayesian Approaches for individualized benefit-risk assessment
Yueqin Zhao1, Ram Tiwari, Jyoti Zalkikar and Siqi Cui
1U.S. Food and Drug Administration
yueqin.zhaof@fda.hhs.gov
Benefit-risk assessment is critical in drug development life cycle. Some benefit-risk measures depend on the probabilities of benefit-risk categories in which the subject-level benefit and risk outcomes are characterized. The existing benefit-risk methods for analyzing the categorical data depend only on the frequencies of mutually exclusive and collectively exhaustive categories that the subjects fall in, and thus ignore the subject-level differences. We propose a Bayesian method for analyzing the subject-level categorical data with multiple visits. A generalized linear model is used to model the subject-level response probability of each category, with respect to a “reference” category, assuming a logit model with subject-level category effects and multiple visit effects. Dirichlet process is used as a prior for the subject-level category effects to catch the similarity among the subject responses. We develop an efficient Markov chain Monte Carlo algorithm for implementing the proposed method, and illustrate the estimation of individual benefit-risk profiles through a simulation study. A clinical trial data is analyzed using the proposed method to assess the subject-level or personalized benefit-risk profile in each arm, and to evaluate the aggregated benefit-risk difference between the treatments at different visits.

Individual Benefit Risk Assessment: Overview and Proposal
Sammy Yuan and Shahrul Mt-Isa
Merck & Co., Inc.
yuans45@gmail.com
In light of patient centered benefit risk project and forthcoming patient focus drug development guidance from regulatory agencies, more focus should be given to individual level benefit and risk assessment. Such individual level BRA, by including patient preference and individual experience data, can help us to choose patient-oriented efficacy or safety endpoints for designing a trial, identify subgroups of patients with more favorable BR profile or with willing to take more risk than other groups, and enable patients who experience greater benefit than risk to access the drug earlier. In this presentation, we will briefly introduce the existing statistical methods for individual BRA, and propose a process to implement the BRA across the life-cycle of drug development.

Keynote lecture III
Keynote lecture III: How good is your selected subgroup?
Xuming He
University of Michigan
Subgroup analysis is frequently used to account for the treatment effect heterogeneity in clinical trials. When a treatment is seen marginally effective for the population of the original study, it is tempting to consider post hoc subgroup identification. When a highly promising subgroup is selected this way, serious questions have to be asked about the potential risks and rewards of the subgroup pursuit. In this talk, we will discuss factors that have direct impacts on the credibility of subgroup pursuit, and then propose a model-free approach to quantify how likely the promise of the selected subgroup is a statistical artifact, and how good the selected subgroup really is. The proposed quantitative analysis of subgroup pursuit can help inform better decisions about any selected subgroup in clinical trials. The talk is based on joint work with Xinzhou Guo.

Session 25: New methods and challenges for analyzing of process data
New methods and challenges for analyzing process data in educational measurement
Zhiliang Ying
Columbia University
yangfengstat+ying@gmail.com
This talk presents some recent work on statistical modeling and analysis of problem-solving process data. Such data arise from modern computer-based tests with items for assessing complex problem solving abilities in technology-rich environments. They consist of sequences of time-stamped events, making it natural to develop counting process-based models that incorporate meaningful latent structures. Examples from the Programme for the International Assessment of Adult Competencies (PIAAC) and the Programme for International Student Assessment (PISA) will be used to illustrate the new developments. Conceptual and technical challenges in the modeling and analysis of process data will also be discussed.

A joint Modeling Framework using responses and response times to track skill acquisition: Model Estimation and Application
Shiyu Wang
University of Georgia
swang44@uga.edu
Response times, which is the amount of time the test taker spends considering and answering each item, has been extensively studied and used in a testing environment as the useful source of information to reflect individual response behavior and item characteristics. In this study, we consider using response time in a learning environment to model students’ learning progress. This type of information, in addition to their responses to the test questions, can be valuable source of information to measure students learning trajectory. We propose a joint modelling framework that use responses
Assessing 21st-century skills via human-human interactions: task design and psychometric challenges

Jiangang Hao
Educational Testing Service
jjhao@ets.org

Collaboration and communication are important 21st-century skills that are crucial for both academic and career successes. Assessing these skills is very challenging. These challenges come from both the design/selection of appropriate tasks and the statistical modeling of the interdependent data. In this talk, I will introduce our work to address these challenges through the ETS Collaborative Science Assessment Prototype (ECSAP) project. I will discuss some preliminary findings as well as some psychometric/statistical challenges that may trigger interests from the statistics community.

Session 26: Biosimilar and bioequivalence

New Proposal for Equivalence Criteria in Bioequivalence Study with Binary Clinical Endpoint

♦ Mengdie Yuan1, Jingyu Luan1 and Hui Sun
1U.S. Food and Drug Administration
mengdie.yuan@fda.hhs.gov

In comparative clinical endpoint studies with a binary outcome, a generic product (TEST) and the Reference Listed Drug (RLD) are considered to be equivalent if the 90% confidence interval for the difference in success or cure rates between the two products is within [-20%, 20%]. This method, however, is not equally sensitive for delimiting a difference between TEST and RLD across the response range. This is particularly problematic when the success rate for RLD is expected to be low. Therefore we propose a two-step method instead. Through simulation, we show that our proposed method can control the passing probability for products with relatively low success rate, and meanwhile maintain high power with reasonable sample size when Test and RLD are actually equivalent.

Sample Size Requirement for Analytical Biosimilarity Assessment

Tianhua Wang
U.S. Food and Drug Administration
tianhua.wang@fda.hhs.gov

FDA recommends a stepwise approach for obtaining the totality-of-the-evidence for assessing biosimilarity between a proposed biosimilar product and its corresponding reference biologic product being considered (FDA, 2017). The stepwise approach starts with analytical studies for assessing similarity in critical quality attributes which are relevant to clinical outcomes. For critical quality attributes that are most relevant to clinical outcomes (Tier 1 CQAs), FDA requires equivalence testing to be performed for similarity assessment based on an equivalence acceptance criteria. In practice, the number of Tier 1 CQAs might be greater than one and should be no more than four. The number of biosimilar lots is often recommended to be no less than 10 and the ratio between the reference product sample size and biosimilar product sample size is recommended within the range from 2/3 to 3/2. Accordingly, we derive the analytical formulas for the power function, and propose the sample size calculation methods using the analytical power functions, based on FDA’s practical recommendations for the equivalence testing currently used in analytical biosimilar assessment.

Bioequivalence and Non-Inferiority Testing in The Presence of Subject-By-Treatment Interaction

Elena Rantou
U.S. Food and Drug Administration
erantou@fda.hhs.gov

Assessment of equivalence between a reference listed drug (RLD) and a generic (test) product, is usually carried using a two-one sided (TOST) confidence interval. The average bioequivalence approach implemented with this confidence interval, is not always applicable because of certain issues that arise from the nature of the data sets. Such issues include the unusually high within-subject and between-subject variability, the limited sample size and the eventually, the presence of subject-by-treatment interaction. In the case of locally-acting dermatological products, the inherent variability induced from the skin, as well as, the differences in skin permeability among subjects, are often responsible for the presence of subject-by-treatment interaction. Different examples will be discussed, in regards to the appropriate endpoint, appropriate for each problem. In all these cases, the statistical tests should be sensitive enough to detect meaningful differences. Additionally, these tests should not be over-sensitive as this would sometimes result to rejecting good generic products. Taking all these considerations into account, he most appropriate test should be chosen in such a way that achieves a high level of statistical power and maintains a low probability of type-I error.

Modified Wald Test for Reference Scaled Equivalence Assessment of Analytical Biosimilarity

♦ Yu-Ting Weng1, Yi Tsong2, Meiyu Shen1 and Chao Wang1
1U.S. Food and Drug Administration
Yu-Ting.Weng@fda.hhs.gov

For the reference scaled equivalence hypothesis, Chen et al. (2017) proposed to use the Wald test with Constrained Maximum Likelihood Estimate (CMLE) of the standard error to improve the efficiency when the number of lots for both test and reference products is small and variances are unequal. However, by using the Wald test with CMLE standard error (Chen et al., 2017), simulations show that the type I error rate is below the nominal significance level. Weng et al. (2018) proposed the Modified Wald test with CMLE standard error by replacing the maximum likelihood estimate of reference standard deviation with the sample estimate (MWCMLE), resulting in further improvement of type I error rate and power over the tests proposed in Chen et al. (2017). In this presentation, we further compare the proposed method to the exact-based method (Dong et al., 2017a) and the Generalized Pivotal Quantity (GPQ) method (Weerahandi, 1993) with equal or unequal variance ratios or equal or unequal number of lots for both products. The simulations show that the proposed MWCMLE method outperforms the other two methods in type I error rate control and power improvement.

Session 27: Clinical trial design and statistical methods for cancer studies

A novel randomized Phase 2 trial design that constrains sample size by requiring a sufficient experimental response rate

♦ Brian Egleston, Samuel Litvin, Stan Basickes and Eric Ross
One difficulty with using one-sample Phase 2 designs, such as the Simon 2-stage design, is that a historical response rate might be difficult to specify. As a result, clinicians have been advocating in the literature that randomized Phase 2 designs be used instead. A further difficulty in using randomized designs in the Phase 2 setting is that the sample size required is often large. In 2017, Litwin et al. proposed a novel two-sample design that is a hybrid of the Simon one sample design and the traditional randomized design. In order to declare the trial a success, the experimental response rate has to be large enough (i.e. the one-sample component) and the treatment effect has to be large enough (i.e. the randomized component). Consider a two arm trial. Let E represent a random variable of the number of responses in the experimental arm and \( C \) represent the number of responses in the control arm. The criteria for declaring the study a success could be that \( E > e \) and \( E - C > d \). In a single stage design, the probability of declaring the study a success (i.e. power) would hence be \( P(E > e)P(E - C > d|E > e) \) under a presumed hypothesis. One can optimize the design and minimize the sample size by judiciously setting both \( P(E > e) \) and \( P(E - C > d|E > e) \). That is, choosing both a desirable new treatment response rate and a promising preliminary treatment effect can provide a sensible study design with a reasonable sample size. The approach has been extended to a two-stage setting with early stopping.

**Sparse Minimum Discrepancy Approach to Sufficient Dimension Reduction with Simultaneous Variable Selection in Ultra-high Dimension**

†Wei Qian\(^1\), Shanshan Ding\(^1\) and Dennis Cook\(^2\)

\(^1\)University of Delaware

\(^2\)University of Minnesota

weiqian@udel.edu

Sufficient dimension reduction (SDR) is known to be a powerful tool for achieving data reduction and data visualization in regression and classification problems. In this work, we study ultrahigh-dimensional SDR problems and propose solutions under a unified minimum discrepancy approach with regularization. When \( p \) grows exponentially with \( n \), consistency results in both central subspace estimation and variable selection are established simultaneously for important SDR methods. Special sparse structures of large predictor or error covariance are also considered for potentially better performance. In addition, the proposed approach is equipped with a new algorithm to efficiently solve the regularized objective functions without the need to invert a large covariance matrix. Promising applications of our proposal are demonstrated through simulations and real data analysis on biomedical studies.

**Optimal timing of futility interim analyses**

†Dong Xi, Paul Gallo and David Ohlssen

Novartis
dong.xi@novartis.com

Futility analyses provide a mechanism to stop a trial early because of low likelihood to achieve its efficacy objective. They are usually motivated by ethical and economic purposes, so that stopping a trial with poor efficacy could save patients and resources for other promising trials. There are various methods to address futility analyses in the literature but most focus on equally spaced interim looks. We consider a constrained optimization framework where the timing and the futility boundary are decided jointly to balance the risks between stopping trials which should continue, and continuing trials which should stop. The average sample size is used as a key parameter, which is evaluated under different degrees of power loss. Alternative objective functions and constraints are compared to assess the operating characteristics of the optimal futility scheme. Numerical results for single and multiple futility looks are provided.

**Hierarchical testing in a group sequential trial when the primary endpoint data become available earlier than the secondary endpoint data**

Jiangtao Gou

Fox Chase Cancer Center

jiangtao.gou@fccc.edu

We propose a generalized framework for critical boundary refinement when conducting hierarchical hypothesis test in a clinical trial involving multiple interim stages. When the hypothesis test follows the stagewise hierarchical rule or the partially hierarchical rule, we provide an improvement on the secondary boundary. This refinement can boost the power to reject the secondary hypothesis.

**Session 28: Advance in Statistical Methods for Large and Complex Data**

A Nonconvex approach to Sparse Generalized Eigenvalue Problem with Applications to Multivariate Statistics

†Kean Ming Tan\(^1\), Zhaoran Wang, Han Liu and Tong Zhang

\(^1\)University of Minnesota

ktan@umn.edu

Sparse generalized eigenvalue problem (GEP) plays a pivotal role in a large family of high-dimensional learning tasks, including sparse Fisher’s discriminant analysis, canonical correlation analysis, and sufficient dimension reduction. Sparse GEP involves solving a nonconvex optimization problem. Most of the existing methods and theory in the context of specific statistical models that are special cases of the GEP require restrictive structural assumptions on the input matrices. In this paper, we propose a two-stage computational framework to solve the sparse GEP. At the first stage, we solve a convex relaxation of the sparse GEP. Taking the solution as an initial value, we then exploit a nonconvex optimization perspective and propose the truncated Rayleigh flow method (Rifle) to estimate the leading generalized eigenvector and show that it converges linearly to a solution with the optimal statistical rate of convergence. Theoretically, our method significantly improves upon the existing literature by eliminating structural assumptions on the input matrices. To achieve this, our analysis involves two key ingredients: (i) a new analysis of the gradient based method on nonconvex objective functions, as well as (ii) a fine-grained characterization of the evolution of sparsity patterns along the solution path. Thorough numerical studies are provided to validate the theoretical results.

Online experimental design for automated high-throughput neural circuit mapping

Shizhe Chen

Columbia University

shizhe.chen@gmail.com

Recent development in neuroscience allows for the recording and manipulation of neural activities at cellular resolution in living animals. We consider the task of learning the physiological connections among neurons (i.e., synaptic connections) in vivo. However, experiments on large volumes of densely-packed neurons (e.g. cortical excitatory neurons) with two-photon optogenetics has proven challenging because of two main problems: 1) stimulation sensitivity and resolution varies across cells due to differential opsin expression and intrinsic excitability, making the precise localization
of connected neurons difficult, and 2) experimental time is severely limited compared to the number of potential connections to map. We present a method which overcomes these limitations using statistical modeling and online experimental design. To infer posterior distributions for the probability of synaptic transmission and opsin expression level of potentially connected neurons, we fit a model with three main components: a neural response model which predicts presynaptic spike rates given the power and location of stimulation targets, a connectivity model which filters presynaptic spike rates into a postsynaptic event rate, and a postsynaptic model which converts the postsynaptic event rate into electrical activity.

To improve efficiency, we implement a closed-loop parallel computational system which designs batches of stimulation targets online based on fast stochastic variational inference of these posteriors. Our experimental system allows us to collect data at 20 trials per second for a large portion of experimental time while analyzing data in the background. We demonstrate the efficacy of our method in vitro and in vivo by learning connectivity in mouse cortex. This is joint work with Ben Shababo, Xinyi Deng, Johannes Friedrich, Hillel Adesnik, and Liam Paninski.

Concentration inequalities for time series
Fang Han
University of Washington
fanghan@uw.edu

This talk introduces new exponential inequalities of (non)degenerate U-statistics and random matrices for weakly dependent data under geometric phi-, alpha-, and tau-mixing conditions.

Statistical Inference for Online Learning and Stochastic Approximation via Hierarchical Incremental Gradient Descent
*Weijie Su*¹ and Yuancheng Zhu
¹University of Pennsylvania
suw@wharton.upenn.edu

Stochastic gradient descent (SGD) is an immensely popular approach for online learning in settings where data arrives in a stream or data sizes are very large. However, despite an ever-increasing volume of work on SGD, much less is known about the statistical inferential properties of SGD-based predictions. Taking a fully inferential viewpoint, this talk introduces a novel procedure termed HiGrad to conduct statistical inference for online learning, without incurring additional computational cost compared with SGD. The HiGrad procedure begins by performing SGD updates for a while and then splits the single thread into several threads, and this procedure hierarchically operates in this fashion along each thread. With predictions provided by multiple threads in place, a t-based confidence interval is constructed by decorrelating predictions using covariance structures given by the Ruppert–Polyak averaging scheme. Under certain regularity conditions, the HiGrad confidence interval is shown to attain asymptotically exact coverage probability. Finally, the performance of HiGrad is evaluated through extensive simulation studies and a real data example. An R package higrad has been developed to implement the method.

Adaptive Estimation in Structured Factor Models with Applications to Overlapping Clustering
Yang Ning
Cornell University
yang.hbs92263@gmail.com

This work introduces a novel estimation method, called LOVE, of the entries and structure of a loading matrix $A$ in a sparse latent factor model $X = AZ + E$, for an observable random vector $X$ in $\mathbb{R}^p$, with correlated unobservable factors $Z \in \mathbb{R}^K$, with $K$ unknown, and independent noise $E$. Each row of $A$ is scaled and sparse. In order to identify the loading matrix $A$, we require the existence of pure variables, which are components of $X$ that are associated, via $A$, with one and only one latent factor. Despite the fact that the number of factors $K$, the number of the pure variables, and their location are all unknown, we only require a mild condition on the covariance matrix of $Z$, and a minimum of only two pure variables per latent factor to show that $A$ is uniquely defined, up to signed permutations. Our proofs for model identifiability are constructive, and lead to our novel estimation method of the number of factors and of the set of pure variables, from a sample of size $n$ of observations on $X$. This is the first step of our LOVE algorithm, which is optimization-free, and has low computational complexity of order $p^2$. The second step of LOVE is an easily implementable linear program that estimates $A$. We prove that the resulting estimator is minimax rate optimal up to logarithmic factors in $p$. The model structure is motivated by the problem of overlapping variable clustering, ubiquitous in data science. We define the population level clusters as groups of those components of $X$ that are associated, via the sparse matrix $A$, with the same unobservable latent factor, and multi-factor association is allowed. Clusters are respectively anchored by the pure variables, and form overlapping sub-groups of the $p$-dimensional random vector $X$. The Latent model approach to OVErlapping clustering is reflected in the name of our algorithm, LOVE. The third step of LOVE estimates the clusters from the support of the columns of the estimated $A$. We further guarantee cluster recovery with zero false positive proportion, and with false negative proportion control. The practical relevance of LOVE is illustrated through the analysis of an mRNA-seq data set, devoted to determining the functional annotation of genes with unknown function.

Session 29: Statistical Issues for Prognostic Biomarker Discovery and Risk Prediction in Cancer Genom

Measuring differential treatment benefit to evaluate predictive cancer biomarkers
*Jaya Satagopan* and Alexia Iasonos
Memorial Sloan Kettering Cancer Center
satagopj@mskcc.org

Clinical and epidemiological studies of anticancer therapies increasingly seek to identify predictive biomarkers to obtain insights into variation in treatment benefit. For time to event endpoints, a predictive biomarker is typically assessed using the interaction between the biomarker and treatment in a proportional hazards model. Interactions are contrasts of summaries of outcomes and depend upon the choice of the outcome scale. In this work, we investigate interaction contrasts under three scales - the natural logarithm of hazard ratio, the natural logarithm of survival probability, and survival probability at a pre-specified time. We illustrate that we can have a non-zero interaction on survival or logarithm of survival probability scales even when there is no interaction on the logarithm of hazard ratio scale. Since survival probabilities have clinically useful interpretation and are easier to convey to patients than hazard ratios, we recommend evaluating a predictive biomarker using survival probabilities. We provide empirical illustration of the three scales of interaction for evaluating a predictive biomarker using reconstructed data from a published melanoma study.

Gaussian process regression for survival time prediction with
A Bayesian Joint Model of Recurrent Events and a Terminal Event

Ming Wang1, Zheng Li2 and Vern M. Chinchilli2
1Pennsylvania State University College of Medicine
2Pennsylvania State University

Recurrent events could be stopped by a terminal event, which commonly occurs in biomedical and clinical studies. In this situation, the non-informative censoring assumption could be failed because of potential dependency between these two event processes, leading to invalidate inference if analyzing recurrent events alone. The joint frailty model is one of the widely used approaches to jointly model these two processes by sharing the same frailty term. One crucial assumption is that recurrent and terminal event processes are conditionally independent given the subject-level frailty; however, this could be violated when the dependency may also depend on time-varying covariates across recurrences. Furthermore, the marginal correlation between two event processes based on traditional frailty modeling has no closed form solution for estimation with vague interpretation. In order to fill these gaps, we propose a novel joint frailty-copula approach to model recurrent events and a terminal event with relaxed assumptions. Metropolis-Hastings within the Gibbs Sampler algorithm is used for parameter estimation. Extensive simulation studies are conducted to evaluate the accuracy, robustness and predictive performance of our proposal. The simulation results show that compared with the joint frailty model, the bias and mean squared error of the proposal is smaller when the conditional independence assumption is violated. Finally, we apply our method into a real example extracted from the MarketScan database to identify potential risk factors and study the association between recurrent strokes and mortality.

Cure Model in Cancer Immunotherapy

Shu Wang1, Chengxing (Cindy) Lu2 and Rong Liu2
1University of Pittsburgh
2Bayer

In oncology drug development, achievement of cure in a fraction of long term survivors has become an increasingly promising trend, especially with the recent emerging development of immunotherapy. However, traditional Cox regression model is facing challenges in various aspects such as potential violations to the proportional hazard assumption. In addition, traditional Cox regression model emphasizes treatment effect on the failure time, rather on the effect on proportion of long term survivors (i.e., cure rate), which could be of primary interest in situations when a drug differentiates itself by long term benefits with large cure rate. Alternative, cure model is a method that does not require proportional hazard assumption.
and incorporates treatment effect estimates on both failure time and
cure rate. In this presentation, we will demonstrate both models
in a real data example (Eastern Cooperative Oncology Group trial
EST 1684) and assess the impact of inappropriate model type (mix-
ture vs non-mixture cure mode), parametric model mis-specification
and insufficient follow-up time on the estimates on both failure time
and cure rate, in 5 possible real-life scenarios. Recommendations on
cure model specification, sufficient follow-up time and comparisons
between Cox regression and cure model under different situations
are provided. Some practical considerations are also discussed.

Missing data sensitivity analysis for time-to-event data in clini-
cal trials

Xuezhou Mao, Chuanping Yu, Li Qi and Lan Yu
Sanofi
xuezhou.mac@sanofi.com

This research addresses the sensitivity analysis for the robustness of
primary conclusion drawn from the data analysis in survival clini-
cal trials, where the endpoints of interest include the time-to-single
event and the time-to-recurrent events. The independent censoring
is a key assumption that most statistical methods for the time-to-
event endpoints rely on. However, in reality, non-ignorable censor-
ing is likely to occur to subjects in a clinical trial due to various
reasons, for instance the lack of efficacy, severe adverse event or
etc.. The independent censoring could be hardly verified using the
observed data, which would affect the validity of the results based
on the primary analysis to the data. Our research addresses the eval-
uation of sensitivity via the tipping point analysis, which requires
no assumption on the missing data mechanism and is realized via
assessing the clinical possibility of the "margin point" that reverses
the primary conclusion. For the time-to-recurrent event endpoint,
we utilize a counting process to model the number of events, pos-
tulate the Andersen-Gill (AG) model for the event time and con-
duct multiple-imputation from the distribution (with delta-adjusted
hazard) of the inter-event time; and, the scenario of time-to-single
event can be addressed as a special case by replacing the AG model
with the Cox proportional hazard model. Simulation studies were
performed to demonstrate and evaluate the proposed tipping point
sensitivity analysis for the time-to-event data with independent and
non-ignorable censoring.

Validity of Real-World Control Arms for Single-Arm Clinical
Trials: An Oncology Case Study

Zhanglin Cui, Yanping Wang and Douglas Faries
Eli Lilly and Company
cui.zhanglin@lilly.com

Background: Single-arm trials (SAT) with historical controls (HC)
are often used in oncology drug development and, at times, for reg-
ulatory decision-making. However, the lack of a concurrent con-
trol arm is a major challenge for interpretation of the results. The
standard HC is a treatment arm from a previous clinical trial with
patient-level data. The primary comparison between SAT and HC
controls for measured confounders typically using covariate re-
gression. In some cases, only summary data (no patient-level informa-
tion) from a HC are available or there may be no appropriate HC
data at all. Even if HC data from a prior clinical trial are available,
there may be concerns with potential bias due to the fact the data
are not concurrent. Thus, interpreting the SAT results is extremely
challenging. This study proposes a new approach using real-world
patient-level data to create a control arm (RWC) for SAT and an
idea to validate the approach. The Approach: The new approach
consists of an internal validity module and an external validity mod-
ule. The internal validity module includes a primary comparison
between SAT and HC, a secondary comparison between SAT and
RWC, and an equivalence test between HC and RWC. Confounding
control methods such as entropy balancing (EB), inverse probabil-
ity weighting, and propensity score matching are applied to achieve
covariate balance before the comparisons. The equivalence test is
essential to address the validity of the use of RWC to aid SAT
decision-making. The underlying hypothesis is that the drug ef-
fect will be the same if the patient populations are the same. Un-
der this hypothesis, if the HC and the RWC are nearly equivalent,
then the secondary comparison between SAT and RWC is validated.
The external validity module uses a randomized controlled trial to
mimic an SAT and an HC in the internal validity module in order
to achieve the generalizability of the validation in larger samples of
patient population. Case Study: An anonymous case study in lung
cancer was used to illustrate the application and validation of the
approach. In the case study, EB was used to find weights for con-
trols in order to balance covariates prior to treatment comparison.
Balance assessment for each covariate was conducted before and af-
ter the EB reweighting using normalized differences. Kaplan-Meier
method and Cox proportional hazards regression were applied to the
rewighted and balanced treatment groups in assessing overall sur-
vival between the compared treatments. Conclusion: The case study
illustrated how to compare SAT to RWC and validate the compar-
sions in the internal and external validity modules. Future research
should use a simulation study to assess the performance of the ap-
proach, assess the need for calibration of the secondary comparison
when the equivalence is not perfectly satisfied, and assess the need
for separate concurrent RWC for the equivalence test and secondary
comparison.

Session 31: Effective Communication

Effective Communication

We communicate every day of our lives, but do we think about the
message(s) we are sending, do we choose our words carefully so as
to clearly convey our message without ambiguity and to the right
audience (without assuming they either don’t know anything about
what we are saying or they already have certain assumption which
in both cases is often not true). Communicate is particularly impor-
tagant for statisticians, who often need to discuss and interpret very
complex information to an audience who "only wants to know the
p-value" or "whether there are more than 10% of Adverse Event X".
Communication training probably exists at every company, and sci-
entific communication training also exists at many, but this session
will go even further not only in terms of how to clearly convey very
complex interpretation of data to an audience, answer questions in
a way that does not offend anyone (i.e., answer objectively using
the data as a reference rather than using our subjective opinions)
and, lastly, to be very clear and anticipate questions from an audi-
cence who may be particularly naive when it comes to data or, on
the other hand, very savvy. The goal of this session is to prepare
the presenter to come away with the skills that will work for any
type of statistical presentation and, of course, make you even more
attractive to upper management and/or future employees.
Session 32: Statistical Process Control

On optimal control chart of change-point detection for dependent observation sequence

Dong Han
Shanghai Jiao Tong University
donghan@sjtu.edu.cn

A general measure is presented to evaluate the performance of a control chart for detecting changes in distribution. We not only develop a method of determining a random dynamic control limits (alarm thresholds) but also construct a large class of optimal control charts for finite dependent observations. It is shown that the optimality of the limiting optimal control charts when the number of observations goes to infinite can include the known results on the optimal control charts for the infinite i.i.d. observation sequence with different pre-change and post-change distributions. Moreover, a unified expression of the minimum value of the generalized out-of-control average run length for finite and infinite observation sequences is given.

Run Length Properties of The EWMA Control Chart With Changing Limits

Gaofeng Lv
Shanghai Jiao Tong University
eaco9922@163.com

The EWMA control chart has been developed and used widely in practice. Normally, the control limit of nearly all often used control charts is constant. Although there are some works on the control charts with variable control limits, the variable control limits are non-random and not depend on the real-time observations. Here, we define an EWMA control chart with a variable control limit which is a function of the observations (EWMA-VCL) and present the properties of the average run length (ARL) of the EWMA-VCL control chart, compare the performance of the EWMA-VCL control chart with other EWMA control charts. The comparative results of both theoretical analysis and numerical simulation show that the EWMA-VCL control chart is best among all the three control charts mentioned above in detecting the mean shift of the observation process.

A Distribution-Free Control Chart for Monitoring High-Dimensional Processes with Individual Observation

♦ Jinyu Fan and Lianjie Shu
University of Macau
jinyu.fan@connect.umac.mo

Distribution-free control charts are useful in high-dimensional processes because the underlying distribution is often non-normal. However, the design of distribution-free control schemes remains challenging, and few studies have addressed the high-dimensional case. In this paper, we develop a distribution-free control chart for high-dimensional individual observation when only a few historical observations are available. The proposed chart is established based on the Minkowski distance of the interested measurements. A self-starting scheme is incorporated into the estimation of the in-control mean to handle sequential monitoring by simultaneously updating the parameter estimation and checking for out-of-control conditions. Dynamic control limits are also applied to ensure that a specified false alarm rate is attained at each step when a small sample size of historical observations is available. In addition, updated formulas are deduced to accelerate the determination of the dynamic control limit. The distribution-free property of the designed chart lies in the sense that the in-control run length distribution is proved to be the geometric distribution and that the control limit does not depend on the underlying distribution. Theoretical and numerical studies both verify that the proposed chart can achieve the desired IC ARL for any distribution and dimension. The simulation results also show that the proposed control chart is powerful in various kinds of situations, especially for moderate to large shifts when compared with the existing high-dimensional two sample tests. One more attractive feature is that the proposed control chart can quickly detect the mean shifts for heavy-tailed and skewed distributions.

A Hybrid Hierarchical Bayesian Model for Spatio-Temporal Surveillance Data

Jian Zou
Worcester Polytechnic Institute
jzou@wpi.edu

Due to the low signal-to-noise ratio and high dimensional structure, spatio-temporal data analysis is challenging. In outbreak detection, the assumptions for control charts, including independence, normality, and stationarity, are often violated in syndromic surveillance data. We develop a novel hybrid hierarchical Bayesian model through the combination of Dirichlet process and particle filters to resolve these issues. We employ a modified adjacency matrix as the observation matrix in the Markovian state-space model. This methodology achieves dimension reduction and computational efficiency, and also enjoys a superior detection performance with the ability to incorporate online updating for data streaming applications. Our dataset is derived from the Indiana Public Health Emergency Surveillance System. It consists of surveillance data on emergency room visits for influenza-like and respiratory illness from 2008 to 2010. We are able to detect the 2009 H1N1 outbreak as well as the seasonal influenza outbreaks. Numerical results show that our DPPF models improve the outbreak detection performance in both simulation and real data analysis.

Session 33: Decision making in early clinical development of new drugs

PKQT Model Selection, Diagnostics and Prediction for One Oncology Drug

Shu Yang
Novartis
shui.yang@novartis.com

Mixed effects models are widely used to explore the relationship between change from baseline QTc and drug concentration. For one oncology drug, three different mixed effects models (linear, Emax, and asymptotic) and a few baseline covariates were explored. A likelihood based cross-validation method was proposed for the model selection. Different model diagnostic metrics (marginal normality, and asymptotic) and a few baseline covariates were explored.

Dose-finding designs in the presence of disease progression

♦ Bin Cheng1, Lucie Biard and Shing Lee
Columbia University
bc21538@cumc.columbia.edu

In traditional dose finding studies, dose limiting toxicity is determined within a fixed time window where DLT is often treated as a binary outcome. For patients in advance stages, disease progression may occur within the toxicity observation window which leads to unmeasurable DLT. The objective of this work is to investigate the
impact of disease progression on the implementation of the current
dose finding methods and compare the performance of several ad
hoc modifications of the existing dose finding methods by extensive
simulation studies. A practical guidance will be provided.

An ROC based approach to interim Go/No-Go decision-making
in clinical development
Lanju Zhang
Abbvie, Inc.
LANJU.ZHANG@ABBVIE.COM

Interim analysis is often conducted to make an early Go/No Go de-
cision based on available data. There are many methods to spec-
ify Go/No Go criterion; for example, conditional power, predictive
power, and predicted confidence interval. The quality of the Go/No
Go criterion must be evaluated so that the risk of either stopping an
effective drug or continuing an ineffective drug is minimized in a
balanced way. In this presentation, we introduce a receiver operat-
ing characteristic curve (ROC) approach that 1) unifies these three
methods; 2) helps to evaluate quality of different Go/No Go crite-
rria specified by these methods; and 3) selects the timing of interim
analysis in a general setting.

Using Bayesian proof-of-concept (PoC) criteria to design ran-
domized Phase II oncology trials stratified by biomarker status
Yong Zhang
Eisai

In the early stage development of a targeted therapy, assumed to be
more effective in a specific subpopulation, it is very challenging to
decide whether clinical trials should be conducted in all patients or
only in the specific subpopulation. We propose to use randomized
Phase II study results to help make that decision in a subsequent
Phase III study. Traditional frequentist designs often use Type I/II
error rates only to determine sample sizes and make decisions based
on p-values. However, critical values needed for statistical signif-
icance are often implicit and a significant p-value does not neces-
sarily imply a clinically meaningful effect size. We propose some
Bayesian PoC criteria with clearly defined effect size, which require
a reasonably high probability that treatment is better than control,
and the estimated hazard ratio reaches a clinically meaningful size.
Frequentist probabilities of making the right decision for various
scenarios are investigated. Simulations show that the proposed de-
design has good frequentist operating characteristics. The new design
has been applied in a new randomized Phase II trial.

Session 34: Statistical inference in data science

Robust Jackknife Empirical Likelihood for Non-smooth U-
structure Equations
Yongli Sang
University of Louisiana at Lafayette
yongli.sang@louisiana.edu

In many applications, parameters of interest are estimated by solv-
ing some non-smooth estimating equations with U-statistic struc-
ture. Jackknife empirical likelihood (JEL) approach can solve this
problem efficiently by reducing the computation complexity of the
empirical likelihood (EL) method. However, as EL, JEL suffers the
sensitivity problem to outliers. In this paper, we propose a robust
jackknife empirical likelihood (RJEL) to tackle the above limitation
of JEL. The proposed RJEL tilts the JEL function by assigning
smaller weights to outliers. The asymptotic of the RJEL statistic is
derived. It converges in distribution to a multiple of a chi-square
random variable. The multiplying constant depends on the weight-
ing scheme.

Empirical likelihood for the bivariate survival function under
univariate censoring
Yichuan Zhao and Haitao Huang
Georgia State University
yichuan@gsu.edu

The bivariate survival function plays an important role in multivari-
ate survival analysis. Using the idea of influence functions, we de-
velop empirical likelihood confidence intervals for the bivariate sur-
vival function in the presence of univariate censoring. It is shown
that the empirical log-likelihood ratio has an asymptotic standard
chi-squared distribution with one degree of freedom. A comprehen-
sive simulation study shows that the proposed method outperforms
both the traditional normal approximation method and the adjusted
empirical likelihood method in most cases. The Diabetic Retinopa-
thy Data are analyzed for illustration of the proposed procedure.

Smooth and non-smooth functionals of the NPMLE under in-
terval censoring
Zhigang Zhang
Memorial Sloan Kettering Cancer Center
zhangle@mskcc.org

In this talk I will present some theoretical advances on the topic of
smooth and non-smooth functionals of the NPMLE of the distri-
bution function when a survival time is interval censored. Both
asymptotic distributions and convergence rates will be examined.
The results are applied to several inference approaches to illustrate
their utility.

Autoregressive Model for matrix valued time series
Han Xiao
Rutgers University
hx498@stat.rutgers.edu

In finance, economics and many other fields, observations in a ma-
trix form are often observed over time. For example, several key
economic indicators are reported in different countries every quar-
ter. Various financial characteristics of many companies are re-
ported over time. Import-export figures among a group of countries
can also be structured in a matrix form. Although it is natural to turn
the matrix observations into a long vector then use standard vector
time series models, it is often the case that the columns and rows
of a matrix represent different sets of information that are closely
interplayed. We propose a novel matrix autoregressivemodel that
maintains and utilizes the matrix structure to achieve greater dimen-
sional reduction as well as easier interpretable results. The model
can be further simplified by a set of reduced rank assumptions. Esti-
mation procedure and its theoretical properties are investigated and
demonstrated with simulated and real examples.

Session 35: Advances in the pediatric drug development
with extrapolation

Utilization of Adult Data in Designing Pediatric Pharmacoki-
netic Studies: How Much is Historical Adult Data Worth?
Chyi-Hung Hsu and Steven Xa
Janssen Research and Development, LLC
chsu3@its.jnj.com

Pediatric drug development is challenging in many aspects, partic-
ularly, a slow and difficult recruitment may contribute to the failure
of a pediatric program. To increase the efficiency of pediatric drug
development, the extrapolation approach, based on adult data and
other information, was proposed by the US Food and Drug Administration (FDA) in 1990’s. The impact of this approach was evaluated by Dunne, et al (2011), and they concluded that extrapolating did streamline pediatric drug development and help increase the number of approvals for pediatric use. On the other hand, the appropriate level of extrapolation, e.g. full or partial, is often determined by the similarity in disease progression, and in response to intervention. And the decision on extrapolation level usually needs to be made and agreed upon in a very early stage, sometimes even before the collection of pediatric data. In this presentation, a Bayesian alternative is investigated, of which the level of extrapolation and information borrowing is based on the concordance of adult and pediatric data. The same methodology is also applicable to the cases of which leveraging information of younger children to neonates. The benefits of extrapolation will be quantified in terms of sample size reductions.

Modeling and simulation for planning and assessing the PK/PD consistency and comparison between pediatric and adult data
*Mei Zhang and Zhaoling Meng
Sanofi
mei.zhang@sanofi.com

Clinical trial modeling and simulation (M&S) works in different pediatric study planning and development paths. Due to the significant time lapse between adult approval and label update with pediatric data in average, as well as the slow progression and high failure rate in pediatric studies, the agency is open to PK/PD extrapolation approach for pediatric study development. Three different paths can lead to pediatric approval that is full extrapolation, partial extrapolation, and no extrapolation or partial extrapolation from adults to kids. We discuss different roles that modeling and simulation approach plays in these paths. For full extrapolation, the PK based modeling and simulation approach is used to select a pediatric dose(s). For partial extrapolation, M&S is utilized to do the PK/PD bridging and consistency check between the corresponding adult and pediatric studies. For the third path, the M&S approach can help to design a pediatric phase III study. Case studies with appropriate M&S approaches are used to illustrate how the modeling and simulation leverage existing data and knowledge to provide quantitative recommendations, and assist in defining the design of pediatric studies.

Borrowing information from adult studies into pediatrics trials: Bayesian approach and a case study
Jingjing Ye
FDA
jingjing.ye@fda.hhs.gov

Studies of drugs in pediatric patients often face many practical challenges, such as low prevalence patient population, low willingness for parents to enroll children on studies, concerns relating to ethical issues and a perception of increased liability of testing drugs in children and many more. Under PREA (Pediatric Research Equity Act) and expanded under the recently enacted FDARA (FDA Reauthorization Act of 2017), sponsors are required to conduct clinical trials in children for safety and effectiveness of their product. Additionally, studies conducted under BPCA (Best Pharmaceuticals for Children Act) receive financial incentives for voluntarily conducting studies in children. Given the many challenges in studies in children and the wealth of information contained in adult studies, Bayesian approaches have been proposed as one possible solution to make pediatric trials more feasible. One methodology approach proposed by Goodman and Sladky in 2005 allows the incorporation of information of adults in the analysis of pediatric clinical trials, which may offer reduced sample size and increased efficiency if the disease is expected to be similar between adults and pediatrics. In this talk, we will discuss the use of Bayesian statistical approaches for incorporating information from adults in the analysis of pediatric studies. In addition, we will also discuss design considerations and the possible sources of information that can be used for borrowing.

We will illustrate the approach using a case study.

Session 36: Model Informed Clinical Drug Development

Model Informed regulatory decisions
Jingyu Yu
U.S. Food and Drug Administration
jingyu.yu@fda.hhs.gov

Model-Informed Drug Development (MIDD) is the development and application of exposure-based, biological, and statistical models derived from preclinical and clinical data sources to inform drug development and regulatory decision making. For example, modeling and simulation have been used in trial design, data interpretation, prediction of drug safety and efficacy, and to explain variability in patient responses to drugs. MIDD has also impacted regulatory decisions on approvability and labeling. The purpose of this presentation is to provide regulatory perspective on MIDD, discuss the current state and future directions.

Connecting the Dots to Inform Clinical Decision Making: PKPD Modeling, Meta-analysis, and Clinical Trial Design Optimization
*Qi Tang, Zhaoling Meng, Lei Ma, Dimple Patel and Christine Veyrat-Follet
Sanofi
qi.tang@sanofi.com

Clinical trial development involves modeling of multiple aspects of drugs and clinical trials, such as pharmacokinetics, pharmacodynamics, efficacy and safety profiles of investigational drugs and competitors, and clinical trial designs. A split and conquer approach was often taken by pharmaceutical companies. While this approach is efficient in addressing each individual modeling problem, however, integrating the results to lead to smart clinical trial design is a big challenge in practice, simply due to lack of generalists familiar with all aspects of different modeling techniques. We demonstrate an example in Sanofi where people with different backgrounds across pharmacometrics, statistics and clinical trial sciences work together to connect the dots and enable evidence based smart Go/No-Go decision makings and clinical trial designs.

Method on the determination of the threshold value for a continuous biomarker
*Jianan Hui and Buxhi Wang
Boehringer Ingelheim Pharmaceuticals, Inc.
jiananhuistat@gmail.com

Advances in molecular technology have enabled the new drug development to shift toward targeted therapy where a subgroup of patients is more likely to benefit from the treatment. In order to identify the subgroup, often a potential predictive biomarker is investigated to dichotomize the population to marker-positive and marker-negative group and optimize the treatment benefit. Under many circumstances, the potential predictive biomarker is measured on a continuous scale. In addition, assuming the biomarker under investigation is truly a predictive biomarker, selection of higher threshold value would result in the reduction of the marker positive size and
the enrollment speed if enrichment is desired. It is then of interest for clinical trial designs to evaluate this threshold value which optimize the balance between marker positive size and efficacy effect size. In particular, we investigate the case with a single continuous biomarker and propose a resampling method based on hypothesis testing to determine the threshold value.

**Session 37: Novel Applications of Statistical Learning Methods in Complex Biomedical Data**

**Clustering Data on the Sphere: State of the Art and a Poisson Kernel-Based Algorithm**

*Marianthi Markatou*

University at Buffalo

markatou@buffalo.edu

Many applications of interest involve data that can be analyzed as unit vectors on a d-dimensional sphere. We present a clustering method based on mixtures of Poisson kernel-based densities and study their connection with other distributions, such as von Mises-Fisher, appropriate for the analysis of directional data. We prove convergence of the associated EM-type algorithm, identifiability of the mixture of Poisson kernel-based densities, and we discuss a method to simulate from these distributions. We propose an empirical densities distance plot to estimate the number of clusters in a data set and apply our clustering methodology to a number of data sets. We evaluate performance of our methods via simulation. Our experimental results show that the new model exhibits higher macro-precision and macro-recall than competing methods based on von Mises-Fisher and Watson distributions.

(Joint work with Mojgan Golzy)

**Identification of Optimal Biomarker Combination for Treatment-Selection in the Presence of Missing Data**

*Yu Cheng* and *Xiao-Hua Zhou*

1University of Pittsburgh

yucheng@pitt.edu

2Pennsylvania State University

lebao@psu.edu

The minimal clinically important difference (MCID) is the smallest change in a treatment outcome that an individual patient would identify as important. In the era of precision medicine, it is of particular interest to study both point and interval estimations for the individualized MCID. The motivating example of this work is the ChAMP trial, which is a randomized controlled trial to compare debridement to observation of chondral lesions encountered during partial meniscectomy. In this trial, the primary outcome is the patient reported pain score one year after the surgery and we are interested in estimating the individualized MCID so that the treatment effect can be further studied. In this paper, we formulate this problem in a classification setting where nonconvex minimization technique is needed for the optimization. Furthermore, we develop the Bahadur representation of the individualized MCID so that its confidence interval can be derived. The proposed method is illustrated via comprehensive simulation studies. We also apply our proposed methodology to the ChAMP trial analysis.

**Risk screening for Alzheimer’s Disease progression with Volume under the ROC Surface**

*Yu Cheng*

University of Pittsburgh

yucheng@pitt.edu

This work aims to utilize the patient population of the Alzheimer Disease Research Center (ADRC), and to identify important risk factors and biomarkers for Alzheimer’s Disease (AD) through novel statistical methods. To systematically evaluate all potential risk factors that are collected in the ADRC, we need a unifying metric that does not rely on any model assumptions. We utilize a recently developed methodology by our group, which explicitly evaluates the impact of a marker on multiple competing events simultaneously without any model assumption. This novel method is then used as a screening tool to rank the importance of all potential risk factors in ADRC.

**Point and Interval Estimations for Individualized MCID**

*Jiwei Zhao*

State University of New York at Buffalo

jiwei2012zhao@gmail.com

The minimal clinically important difference (MCID) is the smallest change in a treatment outcome that an individual patient would identify as important. In the era of precision medicine, it is of particular interest to study both point and interval estimations for the individualized MCID. The motivating example of this work is the ChAMP trial, which is a randomized controlled trial to compare debridement to observation of chondral lesions encountered during partial meniscectomy. In this trial, the primary outcome is the patient reported pain score one year after the surgery and we are interested in estimating the individualized MCID so that the treatment effect can be further studied. In this paper, we formulate this problem in a classification setting where nonconvex minimization technique is needed for the optimization. Furthermore, we develop the Bahadur representation of the individualized MCID so that its confidence interval can be derived. The proposed method is illustrated via comprehensive simulation studies. We also apply our proposed methodology to the ChAMP trial analysis.

**Session 38: Statistical Models for Understanding HIV/AIDS Epidemics**

**Trial Designs for Evaluating Combination HIV Prevention Approaches.**

*Sayan Dasgupta*, *Ying Qing Chen*, *Lili Peng*, *Yixin Wang* and *Thomas Fleming*

1Fred Hutchinson Cancer Research Center

2University of Washington

sdasgupta2@fredhutch.org

Recent development of both biomedical and behavioral interventions provides the potential of maximize their population impact in risk reduction of HIV transmission via combination prevention intervention approaches. However, developing powerful and easy-to-implement clinical trial design(s) to assess the effectiveness of combined biomedical and behavioral interventions has been inadequate. We conduct Monte-Carlo simulation studies via the Cox proportional hazards models for time to incident HIV-transmissions to investigate four leading candidate trial designs: 1) single-factor design, 2) factorial design, 3) actives-versus-control “multi-arm” design, and 4) all-versus-none “kitchen-sink” design, for assessing combination prevention intervention approaches. Their potential public health impact is also investigated. In this paper, we compare the pros and cons among the four designs, and argue that the factorial design is an efficient design particularly suitable for combination prevention intervention approaches when multiple candidate interventions are included.

**Size Estimation of People At High Risk for HIV Infections**

*Le Bao*

Pennsylvania State University

lebao@psu.edu

HIV is often largely concentrated in sub-populations whose behavior puts them at higher risk of contracting and transmitting viruses, such as people who inject drugs, sex workers and men who have
sex with men. It has been recognized that accurately estimating the sizes of the key populations most affected by HIV at regional and national levels is critical for allocating resources and program planning. This is often a difficult task because the populations involved are hard to reach and often hidden, so standard survey and census methods are inadequate. As a result it is often necessary to leverage multiple data sources, each of which may by itself provide limited information. In addition, expert knowledge about the size of the populations can be useful, even if it is not very precise. We develop a Bayesian hierarchical model for estimating the sizes of local and national key HIV affected populations. The model incorporates multiple commonly used data sources including mapping data, surveys, interventions, capture-recapture data and estimates or guestimates from organizations, as well as expert knowledge.

Comparison between HIV routine testing data and sentinel surveillance data

Ben Sheng1, Jeffrey Eaton2, Kimberly Marsh3, Mary Mahy3 and Le Bao1
1Pennsylvania State University
2Imperial College London
3UNAIDS

In recent years, countries have transitioned from conducting sentinel surveillance (SS) to using routine testing (RT) in order to collect data on HIV prevalence. However, many countries shifted from SS to RT without temporal overlap; in the transition period, we cannot distinguish whether a change observed in prevalence is due to a change in underlying prevalence or due to a change in data source. To address this problem, we propose an informative prior distribution for the calibration parameter between SS and RT. We create this informative prior based on maximum likelihood analysis of countries with SS and RT in overlapping years.

Spatial-temporal distribution of newly detected HIV/AIDS cases among women aged ≥15 years old in China, 2010-2016

Fangfang Chen
National Center for AIDS/STD Control and Prevention, China CDC
chenfangfang@chinaaids.cn

Objective: To identify the spatial clustering and its temporal trends among newly detected HIV/AIDS cases among women aged 15 years old and above in China from 2010 to 2016. Methods: Newly detected HIV/AIDS cases among women ≥15 years old in China during 2010-2016 were collected to describe their demographic characteristics and the changing trends, and spatial autocorrelation analysis was conducted at county level by using ArcGIS 10.3. Results: The number of newly detected HIV/AIDS cases among women ≥15 years old was increasing annually during 2010 to 2016, the main transmission route of whom was heterosexual transmission. Both the number and proportion of HIV/AIDS cases among elderly women aged ≥50 years old increased significantly. Results of spatial analysis demonstrated a county-level clustered distribution of HIV/AIDS cases across the country with a rising global Moran’s I value over the years, concentrating on western and southern China, covering 9 provinces/autonomous regions/municipalities (Yunnan, Guangxi, Sichuan, Xinjiang, Guizhou, Guangdong, Chongqing, Henan and Hunan). The temporal trends of hot spots differed by age groups, with the trend of epidemic among younger women aged 15-49 years old shifting towards western border and southern coastal regions, and that of elderly women aged ≥50 years old spreading northward from southwestern regions. Conclusion: Findings indicate an increasing clustering of the HIV epidemic among female adults in China, particularly in western and southern regions. Prevention and intervention strategies targeted to women by age, particularly in regions of increasing HIV epidemic, are urgently needed.
Statistical mediation is a causal chain among three or more variables; the independent variable influences the dependent variable directly as well as indirectly (through the mediators). Mediation analysis for continuous response variables is well developed in the literature, and it can be shown that the indirect effect equals the total effect minus the direct effect. Similarly, the August 2017 International Conference on Harmonization’s draft E9 (R1) addendum on “estimands” outlines the interplay in clinical trials between investigational treatments, endpoints and so-called “intercurrent events” (IEs). The addendum envisions each treatment affecting the endpoint both directly and through those IEs. The correspondences between mediation analysis and analysis of estimands are, thus, striking, suggesting that mediator analysis offers a mathematical framework for conceptualizing and making inferences about estimands and IEs, ultimately enhancing development and regulatory decisions about the treatments. In this research, we apply some mediation analysis concepts to IEs and estimands. We discuss, for instance, estimating the treatment effects on the endpoint attributable to each IE, and deciding whether a given IE explains a clinically meaningful portion of the total treatment effect. We show how these analyses can guide clinical trial design and conduct, such as the selection of inclusion and exclusion criteria and enhancing the accuracy of estimated relationships between treatment and endpoints. Finally, we illustrate how inferences like these can assist regulatory authorities in enacting treatment policy.

K-subtree Pseudo-likelihood method in Phylogenetic tree model

Weifeng Wang

University of Georgia

weifengwang2012@gmail.com

Because of its sound statistical foundations and high accuracy, maximum likelihood (ML) method is widely used to infer molecular phylogenies. However, the computing time goes to unacceptable as the number of taxa increases, especially when we talk about the tree of life which will cover millions of taxa. Distance based methods are very fast but not accurate. In this paper, we proposed a method called K-subtree pseudo-likelihood(KPL) method to approximate the full likelihood method, which was much faster than ML and more accurate than distance-based methods although a little bit less accurate than ML methods.

Inference on the phylogenetic tree sets

Yan Du

University of Georgia

yd92006@uga.edu

The purpose of this study is to find the asymptotic distributions of the phylogenetic trees and then use them to construct the confidence sets and test hypothesis. The phylogenetic trees are estimated stochastic process recovered from the current characteristics of several species. As other estimators, statisticians are trying to treat them as random variables. This study employs the correspondence between phylogenetic tree and its matrix representations to explore the asymptotic distributions of trees. This research is focused on the tree topology distribution. The results of this research can be used to construct the confidence sets of the true tree topology, opening phylogenetic tree analysis to techniques available in the matrix computations. Several simulations are run to support the result.

Quantile Regression for big data with small memory

 Xi Chen1, Weidong Liu and Yichen Zhang

1New York University

xchen3@stern.nyu.edu

In this talk, we discuss the inference problem of quantile regression for a large sample size n but under a limited memory constraint, where the memory can only store a small batch of data of size m. A popular approach, the naive divide-and-conquer method, only works when n = o(m^2) and is computationally expensive. This talk proposes a novel inference approach and establishes the asymptotic normality result that achieves the same efficiency as the quantile regression estimator computed on all the data. Essentially, our method can allow arbitrarily large sample size n as compared to the memory size m. Our method can also be applied to address the quantile regression under distributed computing environment (e.g., in a large-scale sensor network) or for real-time streaming data.

Communication-Efficient Distributed Statistical Inference

Yun Yang1, Jason Lee2 and Michael Jordan3

1Florida State University

2University of Southern California

3University of California, Berkeley

yyang@stat.fsu.edu

We present a Communication-efficient Surrogate Likelihood (CSL) framework for solving distributed statistical inference problems. CSL provides a communication-efficient surrogate to the global likelihood that can be used for low-dimensional estimation, high-dimensional regularized estimation and Bayesian inference. For low-dimensional estimation, CSL provably improves upon naive averaging schemes and facilitates the construction of confidence intervals. For high-dimensional regularized estimation, CSL leads to a minimax-optimal estimator with controlled communication cost. For Bayesian inference, CSL can be used to form a communication-efficient quasi-posterior distribution that converges to the true posterior. This quasi-posterior procedure significantly improves the computational efficiency of MCMC algorithms even in a non-distributed setting. We present both theoretical analysis and experiments to explore the properties of the CSL approximation.

Distributed estimation of principal eigenspaces

Ziwei Zhu1, Jianqing Fan1, Kaizheng Wang2 and Dong Wang

1Princeton University

ziweiz@princeton.edu

Modern data sets are often decentralized; they are generated and stored in multiple sources across which the communication is constrained by bandwidth or privacy. This talk focuses on distributed estimation of principal eigenspaces of covariance matrices with decentralized data. We introduce and analyze a distributed algorithm that aggregates multiple principal eigenspaces through averaging the corresponding projection matrices. When the data distribution has sign-symmetric innovation, the distributed PCA is proved to be “unbiased” such that its statistical error will converge to zero as the number of data splits grows to infinity. For general distributions, when the number of data splits is not large, this algorithm is shown to achieve the same statistical efficiency as the full-sample oracle. We applied our algorithm to implement distributed partition of traffic network of Manhattan; the distributed procedure delivered similar partition results as the centralized procedure provided that the number of data splits is not large.
Statistical Inference for Model Parameters in Stochastic Gradient Descent

Yichen Zhang\(^1\), Xi Chen\(^1\), Jason Lee\(^2\) and Xin T. Tong\(^3\)
\(^1\)New York University
\(^2\)University of Southern California
\(^3\)National University of Singapore
\(yz2756@nyu.edu\)

The stochastic gradient descent (SGD) algorithm has been widely used in statistical estimation for large-scale data due to its computational and memory efficiency. While most existing works focus on the convergence of the objective function or the error of the obtained solution, we investigate the problem of statistical inference of true model parameters based on SGD when the population loss function is strongly convex and satisfies certain smoothness conditions.

Our main contributions are two-fold. First, in the fixed dimension setup, we propose two consistent estimators of the asymptotic covariance of the average iterate from SGD: (1) a plug-in estimator, and (2) a batch-means estimator, which is computationally more efficient and only uses the iterates from SGD. Both proposed estimators allow us to construct asymptotically exact confidence intervals and hypothesis tests.

Second, for high-dimensional linear regression, using a variant of the SGD algorithm, we construct a debiased estimator of each regression coefficient that is asymptotically normal. This gives a one-pass algorithm for computing both the sparse regression coefficients and confidence intervals, which is computationally attractive and applicable to online data.

Session 42: New developments in microbiome sequencing data modeling and analysis

A two-stage microbial association mapping framework with advanced FDR control

Jiyuan Hu\(^1\), Hyunwook Koh\(^2\), Linchen He\(^2\), Huilin Li\(^2\), Menghan Liu\(^2\) and Martin Blaser\(^2\)
\(^1\)NYU Langone Medical Center
\(^2\)NYU School of Medicine
\(jiyuan.hu@nyumc.org\)

In microbiome studies, it is important to detect taxa which are associated with pathological outcomes at the lowest definable taxonomic rank, such as genus or species. Traditionally, taxa at the target rank are tested for association individually, and then the Benjamini-Hochberg (BH) procedure is applied to control for false discovery rate (FDR). However, this approach neglects the dependence structure among taxa and may lead to conservative results. We propose a two-stage microbial association mapping framework (massMap) which uses prior grouping information from the taxonomic tree to strengthen statistical power at the target rank. MassMap first screens the association of taxonomic groups at a pre-selected higher taxonomic rank using a powerful microbial group test OMiAT. Then it proceeds to test the association for each candidate taxon at the target rank within the significant taxonomic groups identified in the first stage. Hierarchical BH and selected subset testing procedures are evaluated to control the FDR for the two-stage structured tests. Extensive simulations and real data analyses have shown that massMap achieves higher statistical power and detects more taxa.

Optimal Error-in-Variable Compositional Regression for Next-Generation Sequencing Data

Pixed Shi, Anru Zhang and Yuchen Zhou
University of Wisconsin-Madison
\(pixushi@stat.wisc.edu\)

In microbiome and genomic research, regression analysis of next-generation sequencing data has been a crucial tool for identifying microbial taxa or genes that are associated with clinical phenotypes. To eliminate the effect of sequencing depth, log-contrast model is often used where read counts are normalized into compositions and zeros are artificially corrected. To avoid subjective zero correction, we introduce a novel error-in-variable compositional regression model with an easy and efficient estimation procedure. We provide theoretical justification for our method and derive lower bound for the estimation error. Simulation analysis demonstrates the advantage of our method under a variety of configurations. The usage of our procedure is also illustrated through real data analysis.

A Random Effects Model for Large-Scale Multinomial Data

Wei Lin and Jingru Zhang
Peking University
\(weilin@pku.edu.cn\)

Large-scale multinomial data are prevalent in many modern applications, where the data are summarized as multinomial counts with a large number of categories for each subject. Among random effects models that account for heterogeneity and overdispersion in such data, the logistic normal multinomial (LNM) model is attractive for its flexibility to allow for a general dependence structure among the categories, but suffers from computational difficulties in high dimensions due to its intractable likelihood. In this article, we develop a stochastic approximation EM algorithm with Hamiltonian Monte Carlo sampling for fast and scalable maximum likelihood and empirical Bayes estimation in the LNM model. We further propose a covariance regularization method by imposing a condition number constraint and investigate the risk property of the regularized estimator. The advantages of the proposed method are illustrated on simulated and human gut microbiome data.

GLM-based latent variable ordination method

Michael Sohn and Hongzhe Li
University of Pennsylvania
\(msohn@pennmedicine.upenn.edu\)

Distance-based ordination methods, such as the principal coordinate analysis (PCoA), are incapable of distinguishing dispersion effect (i.e., an effect caused by the difference in variability) from location effect (i.e., an effect caused by the difference in mean abundance). In other words, PCoA may falsely display a location effect when there is a significant dispersion effect. To resolve this potential problem, we propose, as an ordination method, a zero-inflated quasi-Poisson factor model whose estimated factor loadings can be used to display the similarity of samples.

Session 43: Flexible modeling and inference for complex data

Stagewise Co-Sparse Low-Rank Matrix Decomposition

Kun Chen
University of Connecticut
\(kun.chen@uconn.edu\)

We consider the problem of computing a sparse generalized singular value decomposition (SVD) of a matrix. This co-sparse and low-rank structure is very appealing, particularly in high-dimensional multivariate regression, as it implies that the outcomes are related to the features through a few distinct pathways, each of which may only involve a subset of the outcomes and the features. However,
many existing computational methods involve repeated SVD operations and/or orthogonality constraints, rendering them unsuitable for large-scale problems. Motivated by the power method for computing the SVD, we take a two-step approach to tackle the problem. First, we propose divide-and-conquer strategies to simplify the problem into a set of co-sparse unit-rank regression problems. Then for the resulting unit-rank regression, we innovate a stagewise estimation procedure to efficiently trace out its entire solution path. We show that the new algorithm is able to balance between computational efficiency and statistical accuracy, and as the step size goes to zero, the stagewise solution paths converge exactly to the paths of the corresponding rank-constrained penalized regression. Further extensions to generalized eigenvalue problems and tensor decomposition will be discussed.

On Scalable Inference with Stochastic Gradient Descent

Yixin Fang1, Jinfeng Xu2 and Yang Lei2
1New Jersey Institute of Technology
2Hong Kong University

In many applications involving large dataset or online updating, stochastic gradient descent (SGD) provides a scalable way to compute parameter estimates and has gained increasing popularity due to its numerical convenience and memory efficiency. While the asymptotic properties of SGD-based estimators have been established decades ago, statistical inference such as interval estimation remains much unexplored. The traditional resampling method such as the bootstrap is not computationally feasible since it requires to repeatedly draw independent samples from the entire dataset. The plug-in method is not applicable when there are no explicit formulas for the covariance matrix of the estimator. In this paper, we propose a scalable inferential procedure for stochastic gradient descent, which, upon the arrival of each observation, updates the SGD estimate as well as a large number of randomly perturbed SGD estimates. The proposed method is easy to implement in practice. We establish its theoretical properties for a general class of models that includes generalized linear models and quantile regression models as special cases. The finite-sample performance and numerical utility is evaluated by simulation studies and two real data applications.

Conditional regression based on a multivariate zero-inflated logistic model for human microbiome data

Zhigang Li1 and Hongche Lee2
1Geisel School of Medicine at Dartmouth College
2University of Pennsylvania

Massive high dimensional human microbiome data is commonly seen in molecular epidemiology research and have substantially increased in complexity to address critical health concerns due to complex data structure. Analysis challenges arise from compositional, phylogenetically hierarchical, sparse and high dimensional structure of microbiome data. Compositional structure could induce spurious relationships due to the linear dependence between compositional components. In addition, the hierarchical structure of microbiome data from the phylogenetic tree generates dependence at the hierarchical levels which poses a further modeling challenge. Furthermore, the sparsity of microbiome data due to excessive zero sequencing reads for microbial taxa remains an unresolved issue in the literature. Coupled with the high dimensional feature, microbiome data raises great challenging problems in the field of media data analysis. We will develop a zero-inflated logistic normal model to address these issues. A simulation study will show the performance of the approach and a real study example will be included as well.

Spillover effect of information arrivals in security trading

Haipeng Xing and Zhenning Wang
SUNY, Stony Brook
haipeng.xing@stonybrook.edu

In the standard bivariate mixture distribution model for a single security market, the flow of information arrivals determines the dynamics of stock price volatility and trading volume. Due to the contemporaneous and cross-sectional dependence of information arrivals among several securities, a spillover effect might occur. To model such effect, we extend the mixture distribution hypothesis to the case of multiple securities, and use a latent vector autoregressive process to characterize information arrivals. We further use a stochastic approximation algorithm with Markov chain Monte Carlo method to estimate the dynamics of information arrivals. We apply our models and inference procedure to analyze markets of multiple securities and show the spillover effects in them are significant.

Session 44: Statistical solutions to practical issues in educational testing

Best Linear Prediction of Test True Scores
Lili Yao
Educational Testing Service
lyao@ets.org

In many educational tests which involve constructed responses, a traditional test score is obtained through holistic scoring by human raters. However, human raters use subjective judgment to score a particular response to an open-ended question. Hence, the scoring accuracy for a test composed of constructed response items is imperfect because it depends on fallible human judgments. In applications of classical test theory, a true score can be predicted both by directly-related test results and by related ancillary test information. Given the criteria of minimizing mean square error, best linear prediction (BLP) has been applied to predict the true score based on the available information. This talk aims to introduce the application of BLP in predicting test true scores. The methodology is illustrated using data from an international language assessment.

Modeling Writing Processes Using Keystroke Logs
Hongwen Guo, Mo Zhang, Paul Deane and Randy Bennett
Educational Testing Service
hguo@ets.org

In a writing process, four sub-processes are assumed (Hayes, 2012): proposer (generating ideas and planning), translator and transcriber (translating ideas into words and sentences and writing them out), and evaluator (monitoring progress, editing and reviewing). A major attraction of keystroke logging (KL) is that the moment-by-moment process of text creation is recorded unobtrusively, allowing the text production process to be analyzed more naturally. It They provides an opportunity to analyze writer’s writing states in the process: long pause (proposer), editing (evaluation), and text producing (transcribing). Analyses based on KL can help researchers to understand the strategies that different writers execute during composition, providing more evidence of writing beyond the final production. Our research focuses on how to use stochastic processes to model writing processes in terms of writing states and duration time and how subgroups of writers are different in their writing processes. For this purpose, we developed three steps: first,
Using keystroke timing data and features, we automatically classified keystroke events into a sequence of those three writing states; second, we fit Markov models and Semi-Markov models to the sequences of states and duration time and evaluated model fit; third, we used subgroups matched on essay scores to evaluate their process differences for subgroups. Results will be discussed in the presentation.

**Latent Regression Model for Group Score Estimation: Model Selection and Model Reduction**

Yue Jia  
**Educational Testing Service [Yjieets.org]**

Educational survey assessments are conducted periodically within and across countries to monitor educational progress in student population over years. For instance, in the United States, the survey assessment named National Assessment of Educational Progress (NAEP) has been used to measure students’ progress since 1960th. A common design feature in educational survey assessments is matrix sampling of test questions, that is, students only receive a relatively small portion of the test questions, and common methods for student proficiency estimation do not apply. Instead, marginal likelihood estimation methodologies have been developed through which student answers to test questions are aggregated across student groups using a latent regression model based on Item Response Theory (IRT) models. The independent variables used for these models are often several hundreds of student group indicators obtained through school records and student, teacher, and school questionnaire that are being administered concurrently with the test questions. In this presentation, we will describe the recent development in model selection and model reduction on IRT latent regression as well as their relationship to population proficiency estimates.

**Statistical Procedures for Detecting Large-Scale Unusual Similarity in Responses**

Yi-Hsuan Lee  
**Educational Testing Service [ylee@ets.org]**

In the test-security literature, a common question is whether a specific response pattern occurs unusually often within a group of test takers. This question has been addressed by statistical methods designed for three types of testing irregularities: pairwise answer copying, item preknowledge, and test collusion due to large-scale sharing of test materials/answers to subsets of items before the test administration. In many current tests, modern communication techniques can permit large numbers of test takers to share keys, or common response patterns, to the entire test. This special type of test collusion is termed key sharing. Test takers may communicate in real time while taking a test, or they may receive an answer key from a source who has already had access to the examination. In general, test takers sharing the keys are not limited to the same test location and may have no connection except to the same key sources and keys. In this presentation, I will discuss existing statistical procedures that aim to identify test takers who exhibit unusual similarity in responses to the entire test in groups of any size greater than one.

**Session 45: New Topics in Statistical Learning**

**Individualized Treatment Rules with Reject and Refine Options**

Xingye Qiao\(^1\) and Haomiao Meng  
\(^1\) Binghamton University  

qiao@math.binghamton.edu

One of the goals of the precision medicine research is an optimal individualized treatment rule (ITR), which can select an optimal treatment according to each patient’s specific characteristics. Recently, Outcome Weighted Learning (OWL) has been proposed to estimate ITR in various settings. In practice, an optimal set of multiple treatments may be desired as the output of an ITR, since for certain patients, multiple treatments may have non-distinguishable optimal outcomes. In this case, health care providers may take secondary factors into consideration in the decision process, such as the health care expense, the invasiveness of the treatment, and the life quality of the patient. This motivates a new type of ITR with reject and refine options. We show that this is naturally related to multiclass classification with reject and refine options. We first characterize the optimal set of treatments for each patient, then propose two methods based on the OWL framework to estimate the new ITR. Theoretically, we show Fisher consistency of these methods and obtain an upper bound for the excess risk. A simulation study and a real data example have shown the usefulness of these new learning tools. This is a joint work with Haomiao Meng.

**High-dimensional cost-constrained regression via non-convex optimization**

Guo Yu\(^1\), Haoda Fu\(^2\) and Yufeng Liu  
\(^1\) State University of New York at Buffalo  
\(^2\) Eli Lilly and Company  

guanyu@buffalo.edu

In modern predictive modeling process, budget constraints become a very important consideration due to the high cost of collecting data using new techniques such as brain imaging and DNA sequencing. This motivates us to develop new and efficient high-dimensional cost-constrained predictive modeling methods. In this paper, to address this challenge, we first study a new non-convex high-dimensional cost-constrained linear regression problem, that is, we aim to find the cost-constrained regression model with the smallest expected prediction error among all models satisfying a budget constraint. The non-convex budget constraint makes this problem NP-hard. In order to estimate the regression coefficient vector of the cost-constrained regression model, we propose a new discrete extension of recent first-order continuous optimization methods. In particular, our method delivers a series of estimates of the regression coefficient vector by solving a sequence of 0-1 knapsack problems that can be addressed by many existing algorithms such as dynamic programming efficiently. Next, we show some extensions of our proposed method for statistical learning problems using loss functions with Lipschitz continuous gradient. It can be also extended to problems with groups of variables or multiple constraints. Theoretically, we prove that the series of the estimates generated by our iterative algorithm converge to a first-order stationary point, which can be a globally optimal solution to the non-convex high-dimensional cost-constrained regression problem. Computationally, our numerical studies show that the proposed method can solve problems of fairly high dimensions and has promising estimation, prediction, and model selection performance.

**Likelihood Ratio Test for Stochastic Block Models with Bounded Degrees**

Zuofeng Shang  
Indiana University-Purdue University Indianapolis  

shangzf@iu.edu

A fundamental problem in network data analysis is to test whether a network contains statistical significant communities. We study
this problem in the stochastic block model context by testing $H_0$: Erdos-Renyi model vs. $H_1$: stochastic block model. This problem serves as the foundation for many other problems including the testing-based methods for determining the number of communities and community detection. Existing work has been focusing on growing-degree regime while leaving the bounded-degree case untreated. Here, we propose a likelihood ratio type procedure based on regularization to test stochastic block models with bounded degrees. We derive the limiting distributions as power Poisson laws under both null and alternative hypotheses, based on which the limiting power of the test is carefully analyzed. The joint impact of signal-to-noise ratio and the number of hidden communities on the asymptotic results is also unveiled. The proposed procedures are examined by both simulated and real-world network datasets. Our proofs depend on the contiguity theory for random regular graphs developed by Janson (1995).

A Big Data Linear Regression Via A-optimal Subsampling
Sean Peng and Fei Tan
Indiana University-Purdue University Indianapolis
hanxpeng@iupui.edu

Consider fast approximating the least squares estimator efficiently in a big data linear regression. Given an arbitrary sampling distribution, we use it to take a subsample and construct the approximating subsampling estimator. We derive the expansions for the estimator, its bias, and variance covariance matrix, and establish asymptotic normality under suitable conditions. By minimizing the trace of certain variance covariance matrix over distributions on the data points, we obtain numerous optimal sampling distributions. These include the A-optimal sampling distributions for the subsampling estimator to approximate the LSE, which are given by the normalized square roots of the diagonal entries of the matrix $H_2 = X(X^TX)^{-2}X^T$ or the better A-optimizing matrices $H_2$ and $H_2$. The latter, in particular, are more effective in extracting information than the uniform and the leverage score distribution defined as the normalized diagonal entries of the hat matrix $H = X(X^TX)^{-1}X$. In fact, we use two optimal sampling criteria to derive the leverage scores, both suggesting an ineffective connection of the scores with the approximation to the LSE. We demonstrate that the A-optimal distributions have the same running time as that of the full data LSE, and propose a fast algorithm to compute them. The algorithm possesses the advantages of being implementable by parallel computing and fast updatable for new observations, and having much faster running time than that of the LSE. A data truncation method is introduced. Asymptotic normality is established under various specific sampling distributions. Conditions for the uniform to be A-optimal are given, different sampling distributions compared, and exponential error bounds provided. Our results contribute as well to the weighted bootstrap estimation theory. Our simulations indicate that the A-optimal sampling yielded substantially smaller mean squared errors than the uniform and the leverage scores sampling.

Session 46: Futility analysis and interim decision making

Extending a futility rationale to multiple endpoint settings
Paul Gallo and Dong Xi
Novartis
paul.gallo@novartis.com

Various metrics have been discussed in the literature addressing when a clinical trial might be stopped prematurely on the basis of weak interim results. These include conditional power (under various effect size assumptions), predictive probability, beta spending functions, and others. In multi-look schemes it can be challenging to find criteria that are easily described, yet perform well at the various timepoints at which they might be utilized within a given trial. Understandably most of the literature focuses on criteria for a study’s single main endpoint (even though the reality of when a study should stop is usually more complicated, as a Data Monitoring Committee should factor in other relevant information as well). Studies with multiple endpoints raise additional challenges for determining futility thresholds with desirable properties, as it seems imperative to base actions on consideration of those objectives jointly. This adds additional complexity to determining sensible thresholds compared to familiar single-endpoint situations. This applies in different manners to multiple primary endpoint designs where success can be achieved if a single endpoint reaches significance (with appropriate multiplicity adjustment), as opposed to multiple co-primary endpoint designs where all endpoints must succeed. We should not necessarily expect signals across different endpoints to be highly concordant, and thus need to think carefully about outcome patterns that could really justify stopping. In multiple-endpoint studies, it’s not uncommon that some objectives are over-powered (i.e., to achieve sufficient power for others), so familiar thresholds that may be sensible in other circumstances may not behave as expected. Additionally: different objectives may have different degrees of importance; there may be differing levels of prior belief or evidence regarding favorable outcomes; information may accrue at different rates; correlation among endpoints may be relevant to account for; there may be different ethical implications of poorly-trending results for different endpoints. Based on such considerations, we discuss how we might extend single-endpoint futility rationales to this more complex setting.

Interim Futility Analysis for Longitudinal Data With Adaptive Timing and Error Rate Preservation
G. Cai
GlaxoSmithKline
genqian.z.cai@gsk.com

There are many clinical trials where longitudinal endpoints are used and the primary endpoint is quite often either based on the rate of change or change from baseline at a specific long-term follow-up time point. When such trials are monitored, it is possible that interim futility analyses will be planned such that the trials can be terminated early if the treatment does not induce any benefit to the patients. For such trials, subjects with incomplete follow-up pose challenges in the timing, analysis, and decision making at the interim futility look. We propose an efficient interim futility analysis based on the slope of a linear regression, which incorporates all the data available at the interim analysis. Our approach has the added advantage of providing a data-driven decision on triggering the interim analysis when sufficient information has been collected such that the desired properties for the established futility rule are guaranteed. The construction of interim futility rules and the timing of the interim analysis are discussed and the method is illustrated with an example involving a placebo-controlled comparison of longitudinal proteinuria measurements. Supplementary materials for this article are available online.

Statistical Consideration when Combining Two Independent Trials in Interim Futility Analysis
Dooti Roy, Qiqi Deng, Ming-Hui Chen, and Ying Ying Zhang
Boehringer Ingelheim Pharmaceuticals, Inc.

Abstracts

2018 ICSA Applied Statistics Symposium, New Brunswick, NJ, June 14-17
Traditionally, statistical methods for futility analysis are developed based on a single study. To establish a drug’s effectiveness, usually at least two adequate and well-controlled studies need to demonstrate convincing evidence on its own. Therefore in a standard clinical development program in chronic diseases, two independent studies are generally conducted for drug registration. This paper proposes a statistical method to combine interim data from two independent and similar studies for interim futility analysis and shows that conditional power approach based on combined interim data has better operating characteristics compared to the approach based on single-trial interim data, even with small to moderate heterogeneity on the treatment effects between the two studies.

Futility and the DSMB: An Unholy Alliance?
Janet Wittes
Statistics Collaborative, Inc.

Many Data Safety Monitoring Boards (DSMBs) are charged with reviewing not only safety in an ongoing trial, but also efficacy, often with the view to recommending stopping the trial early if the data are showing insufficient evidence of benefit. In such cases, the DSMB is told to declare the trial “futile”. If the primary endpoint relates to time-to-event and the hazards are not proportional, early data may present a distorted picture of efficacy. Yet many rules for futility fail to account for the possibility of a delayed effect of drug. Examples from lipid-altering trials serve as a cautionary note. When the members of a DSMB and the Sponsor have different views of what constitutes “futility” - (or lack of evidence of efficacy) - the recommendations the DSMB make may be at variance with what the Sponsor would prefer. Many times the futility rule is not stated in a way that is clear enough for the DSMB so that the members agree to a guideline without fully understanding its implications. This talk presents some examples of futility “rules” and discusses their consequences in some real cases.

Session 47: Recent Biomarker Developments with Applications in Brain Research and Disorders

Constructing Concurrent Network of Biomarker Processes Using Dynamical Systems
Ming Sun and Yuanjia Wang
Columbia University
ms47998@columbia.edu

Modeling dynamical system with a large number of components faces the challenge of how to extract structural and regulatory information from noisy observations. In this work, we propose model and method for network construction and parameter estimation using noisy, high-dimensional, multiple time-course data observed from ordinary differential equations (ODEs) systems. Our work is motivated by studies of brain responses to stimuli on patients affected by psychiatric disorders using experiments collecting electroencephalogram (EEG) data. We use single-index model to capture the relationship between the derivatives of each target state variable and all regulatory state variables in the system. To jointly model multiple time-course datasets, we extend our method to double-index model to incorporate subject-level or experimental level features. We propose an integrated index model for ordinary differential equations (IIM-ODE) and a hybrid algorithm for estimation. The network structure sparsity is achieved by introducing regularization in the step of estimating network association parameters.

Adjusting for Covariate Measurement Error in Failure Time Analysis under Competing Risks
Carrie Caswell and Sharon Xie
University of Pennsylvania
caswellc@pennmedicine.upenn.edu

Time-to-event data in the presence of competing risks has been well studied in recent years. One popular approach to this problem is to model the subdistribution of competing risks with assumptions of proportional hazards and covariates measured without error. The estimator resulting from this model does not perform as expected when the covariates are measured with error, which is often the case in biomarker research. We propose a novel method which combines the intuition of the subdistribution model with risk set regression calibration, which corrects for measurement error in Cox regression by recalibrating at each failure time. We perform simulations to assess under which conditions the subdistribution estimator incurs a significant amount of bias in regression coefficients, and demonstrate that our new estimator reduces this bias. We show that our proposed estimator is asymptotically normally distributed and provide a consistent variance estimator. This method is applied to Alzheimer’s Disease Neuroimaging Initiative data, which examine the association between measurement error-prone cerebrospinal fluid biomarkers and risk of conversion to Alzheimer’s disease.

Determine Appropriate Sample Size for a Biomarker Signature Discovery Problem Using Penalized Regression
Xiang Li, Hong Tian and Liang Xiu
Janssen Research and Development, LLC
xli25661@its.jnj.com

Advances in next generation sequencing (NGS) technologies provide exciting opportunities to study disease etiology and pathology. One of the challenges with development of NGS based signatures is to determine appropriate sample sizes. To accommodate the large scale of biomarkers, penalized regression, such as Lasso, is often used to identify biomarkers. The ideal variable selection procedure would search for the best subset of predictors, which is equivalent to imposing an L0-penalty on the regression coefficients, but is known as NP-hard. We proposed an efficient augmented and penalized minimization L0 (APM-L0) to solve L0-penalty variable selection problem. The sample size calculation plays an important role in the study planning but is not well studied, especially for biomarker signature discovery problems using regularization methods. The objective is to use simulation studies to quantify the impact of sample size on the performance of various penalized methods measured by prediction error and selection performance. Our proposed variable selection method can be applied to clinical studies for disease diagnostic and enrichment design.

Session 48: Collaboration Space between Biostatistics and Pharmacometrics

This isn’t peewee soccer - building teamwork between statisticians and pharmacometricians
Bret Musser
Regeneron Pharmaceuticals, Inc.
bret.musser@regeneron.com

Statistics and Pharmacometrics are two disciplines that bring unique perspectives to solving problems in drug development, yet because
of the overlap in quantitative skills there has often been a lot of misunderstanding in how they work together. Effective teamwork requires mutual respect and understanding in what each brings, and this talk will provide a statistician’s perspective on how each discipline contributes to drug development and the opportunities and challenges in building an effective team. A vision for maturing how the disciplines can effectively work together will be shared, drawing from experiences as a statistician and leader who has worked in both a Statistics department and in a Pharmacometrics department.

Synergy between Pharmacometrics and Statistics in Drug Development Decision Making

Liping Zhang
Janssen Research and Development, LLC
lzhanglei@its.jnj.com

Statisticians have long been key players in all areas of drug development and increasingly participating in biological, pharmacological, and real world evidence modelling. Pharmacometricians also use mathematical models for quantitative analysis of interactions between drugs and patients with a focus on population and variability. The nature of the pharmacometric work requires good understanding of statistical disciplines, and pharmacometrics has benefited greatly from the advances in statistical methodology. It is widely agreed that statisticians and pharmacometricians should work together, yet there are still hurdles preventing close collaborations between the two disciplines. This presentation will discuss the opportunities where cross-learning could be helpful and joint efforts be desirable, and propose strategies to overcome some perceived obstacles. Ultimately, strong synergy between pharmacometrics and statistics will lead to better decision making in drug development.

Session 49: Study Designs for Phase I Oncology Dose-Finding Trials

Embracing model-based designs for dose finding trials

Birgit Gaschler-Markefski
Boehringer Ingelheim Pharmaceuticals, Inc.
birgit.gaschler-markefski@boehringer-ingelheim.com

Dose-finding trials are essential to drug development as they establish recommended doses for later-phase testing. We aim to motivate wider use of model-based designs for dose finding, such as the Bayesian logistic regression model (BLRM). We carried out a literature review of dose-finding designs and conducted a survey to identify perceived barriers to their implementation. We describe the benefits of model-based designs (flexibility, superior operating characteristics, extended scope), their current uptake, and existing resources. The most prominent barriers to implementation of a model-based design were lack of suitable training, chief investigators’ preference for algorithm-based designs (e.g., 3\(c^3\>_1; \_c^2\>_3\)), and limited resources for study design before funding. We use a real-world example to illustrate how these barriers can be overcome. There is overwhelming evidence for the benefits of model-based design. Many leading pharmaceutical companies routinely implement model-based designs. Our analysis identified barriers for academic statisticians and clinical academics in mirroring the progress industry has made in trial design. Unified support from funders, regulators, and journal editors could result in more accurate doses for later-phase testing, and increase the efficiency and success of clinical drug development. We give recommendations for increasing the uptake of model-based designs for dose-finding trials in academia.

AAA: Triple-adaptive Bayesian designs for the identification of optimal dose combinations in dual-agent dose-finding trials

Yuan Ji\(^1\) and Jiaying Lyu\(^2\)

\(^1\)University of Chicago
\(^2\)Fudan University
koeraser@gmail.com

We propose a flexible design for the identification of optimal dose combinations in dual-agent dose-finding clinical trials. The design is called AAA, standing for three adaptations: adaptive model selection, adaptive dose insertion, and adaptive cohort division. The adaptations highlight the need and opportunity for innovation for dual-agent dose finding, and are supported by the numerical results presented in the proposed simulation studies. To our knowledge, this is the first design that allows for all three adaptations at the same time. We find that AAA enhances the chance of finding the optimal dose combinations and shortens the trial duration. A clinical trial is being planned to apply the AAA design.

Flexible early-phase design for combination therapies

Nolan Wages
University of Virginia
nwages@virginia.edu

In recent years, investigators have recognized the rigidity of single agent, safety only, traditional designs, rendering them ineffective for conducting modern early-phase clinical trials. Novel approaches are required to address more complex research questions, such as those posed in trials evaluating combinations of targeted therapies or immune-oncology agents. We describe the implementation of a model-based design for identifying an optimal treatment combination, defined by low toxicity and high activity, in several ongoing early-phase trials involving combination therapies. Operating characteristics demonstrate the ability of the method to effectively recommend optimal combinations in a high percentage of trials with reasonable sample sizes. The proposed design is a practical, early-phase method for use with combination therapies. The ability of this design to be applied in a variety of combination studies is discussed.

Time-to-event Bayesian Optimal Interval Design to Accelerate Phase I Trials

Ruitao Lin
University of Texas M.D. Anderson Cancer Center
ruitaolin@gmail.com

Late-onset toxicity is common for novel molecularly targeted agents and immunotherapy. It causes major logistic difficulty for existing adaptive phase I trial designs, which require the observance of toxicity early enough to apply dose escalation rules for new patients. The same logistic difficulty arises when the accrual is rapid. We propose the time-to-event Bayesian optimal interval (TITE-BOIN) design to accelerate phase I trials by allowing for real-time dose assignment decisions for new patients while some enrolled patients’ toxicity data are still pending. Similar to the rolling six design, the TITE-BOIN dose escalation/de-escalation rule can be tabulated before the trial begins, making it transparent and simple to implement, but is more flexible in choosing the target DLT rate and has higher accuracy to identify the MTD. Compared to the more complicated model-based time-to-event continual reassessment method, the TITE-BOIN has comparable accuracy to identify the MTD, but has superior overdose control and is simpler to implement. A numerical study shows that the TITE-BOIN design supports continual accrual, without sacrificing patient safety nor the accuracy of identifying the MTD, and therefore has great potential to accelerate early phase drug development. We have also developed graphical user interface-based software to implement the TITE-BOIN design.
Accuracy, Safety and Reliability of Novel Phase I Trial Designs

Ying Yuan, Lei Nie and Heng Zhou

1 University of Texas MD Anderson Cancer Center
2 U.S. Food and Drug Administration

A number of novel model-based and model-assisted designs have been proposed to find the maximum tolerated dose (MTD) in phase I clinical trials, but their differences and relative pros and cons are not clear to many practitioners. We review three model-based designs, including the continual reassessment method (CRM), dose escalation with overdose control (EWOC), and Bayesian logistic regression model (BLRM), and three model-assisted designs, including the modified toxicity probability interval (mTPI), Bayesian optimal interval (BOIN), and keyboard designs. We conduct numerical studies to assess their accuracy, safety and reliability, and the practical implications of various empirical rules used in some designs, such as skipping a dose and imposing overdose control. Our results show that the CRM outperforms EWOC and BLRM with higher accuracy of identifying the MTD. For the CRM, skipping a dose is not recommended as it substantially increases the chance of overdosing patients, while providing limited gain for identifying the MTD. EWOC and BLRM appear excessively conservative. They are safe, but have relatively poor accuracy of finding the MTD. The BOIN and keyboard designs have similar operating characteristics, outperforming the mTPI, but the BOIN is more intuitive and transparent. The BOIN yields competitive performance comparable to the CRM, but is simpler to implement and free of the issue of irrational dose assignment caused by model misspecification, thereby providing an attractive approach for designing phase I trials.


Session 50: Design and execution of late stage oncology trials in modern cancer drug development

A pitfall in monitoring survival trials under non-proportional hazards

Xiaodong Luo
Sanofi

Testing and estimation in survival trials are usually done by log-rank test and Cox model respectively. This is valid in the presence of proportional hazards between the treatment group and the control group. In the presence of non-proportional hazards, different testing procedures (including weighted log-rank, weighted Kaplan-Meier, combo test) are becoming increasingly acceptable. These test statistics may or may not correspond to a meaningful treatment effect measurement, therefore, different effect measure such as restricted mean survival time (RMST), landmark survival time, weight average of hazard ratios are proposed. However, these test and estimating procedures are often concerned at a certain time point. As such, evaluation at different time points may provide completely different results. The necessity of evaluations at different (earlier) time points includes the interim analyses and the totality of the evidence. In the presence of non-proportional hazards, the testing in different time points may reflect different non-null null hypotheses, which therefore require dedicated multiple comparison procedure to control the potential inflation of type-I error rate. In this talk, I will illustrate this point in more detail.

A Flexible Survival Model for Fitting Time to Event Data in Clinical Trials

Jason Liao and Frank Liu

1 Merck & Co., Inc.

Time-to-event data are common in clinical trials to evaluate survival benefit of a new drug, biological product, or device. The commonly used parametric models including exponential, Weibull, Gompertz, log-logistic, log-normal, are simply not flexible enough to capture complex survival curves observed in clinical and medical research studies. On the other hand, the non-parametric Kaplan Meier (KM) method is very flexible and successful on catching the various shapes in the survival curves but lacks ability in predicting the future events such as the time for certain number of events and the number of events at certain time, and predicting the risk of events (e.g., death) over time beyond the span of the available data from clinical trials. It is obvious that neither the non-parametric KM method nor the current parametric distributions can fulfill the needs in fitting survival curves with the useful characteristics for predicting. In this paper, a full parametric distribution constructed as a mixture of three components of Weibull distribution is explored and recommended to fit the survival data, which is as flexible as KM for the observed data but have the nice features beyond the trial time, such as predicting future events, survival probability, and hazard function.

Multiple Hypothesis Testing and Timing of Analyses in Immuno-oncology Trials

Christine Gause and Keaven Anderson

Merck Research Laboratories

Testing for a pre-specified subgroup in oncology clinical trials in addition to all subjects is increasingly common as it is possible that a portion of the study population may achieve greater benefit from the therapy under study than all subjects. This testing approach is further complicated in immuno-oncology trials with time to event endpoints where a delayed separation of survival curves may occur in one or more subgroups of interest. We consider the problem of testing multiple endpoints in multiple subgroups with multiple interim analyses in immune-oncology trials under the conditions of varying degrees of delayed separation of survival curves. An extension of graphical methods for multiplicity adjustment for group sequential designs by Maurer and Bretz (2013) is applied. Examples using calendar-based as well as information-based spending approaches are considered.

Apply Bayesian Response Adaptive Design to Select Winner from Novel Cancer Treatment

Xiaoling Wu

Celgene

I propose a model-free Bayesian response adaptive design to select the winner between novel cancer treatments. My work is motivated by recent umbrella trials in oncology. But extends by exploring novel treatments on the same patient population, with no known biomarkers established. There are two stages. In stage 1, fixed number of subjects will be randomized to each treatment arm with equal probability. At the end of stage 1, a treatment arm can be dropped due to futility determined by predictive probability of efficacy. At the second stage, each new subject will be dynamically allocated to a treatment arm with the highest predictive probability of response. Treatments that perform well will be increasingly assigned to graduate and win. Conversely, treatments that perform poorly will be
Keynote lecture IV

Keynote lecture IV: Statistical Inference With High-Dimensional Data
Cun-Hui Zhang
Rutgers University

We consider statistical inference in a semi-low-dimensional approach to the analysis of high-dimensional data. The relationship between this semi-low-dimensional approach and regularized estimation of high-dimensional objects is parallel to the more familiar one between semiparametric analysis and nonparametric estimation. Low-dimensional projection methods are used to correct the bias of regularized high-dimensional estimators, leading to efficient point and interval estimation. Bootstrap can be used to carry out simultaneous inference. Only a small fraction of labelled data are needed in a semisupervised setting. Examples include regression and graphical models for continuous and binary data.

Session 51: Challenges and advances in imaging-genetics

Challenges and Progress in Imaging Genetic - Big Data Squared - Studies
Heping Zhang
Yale University

Neuroimaging has been an essential tool for collecting data on the functioning of the brain. High throughput technologies have provided ultra-dense genetic markers to enable us in identifying genetic variants for complex diseases. Only until recently, datasets of reasonably large scale become available that contain both imaging and genetic data. Due to the complexities and high dimensionality in such data, most of the existing datasets are still relatively small in sample sizes but larger datasets are in the horizon. Thus, it is timely and important to develop statistical methods and analytic tools to analyze imaging genetic data - the so-called big data squared. In this talk, I will present the basic technologies, concepts, challenges, and methods related to imaging genetic data. I will use specific data on learning disorders illustrate how to quantify neurobiological risk for learning problems with neuroimaging biomarkers and how to integrate imaging and genetic data in our understanding of cognition and genetic etiologies. This is a joint work with Hongtu Zhu, Xuan Bi, Long Feng, Canhong Wen, and Chintan Mehta. This research is supported in part by NIH grants.

Bayesian nonparametric method for genetic dissection of brain activation region

Jian Kang1, Zhuxuan Jin2 and Tianwei Yu2

1University of Michigan
2Emory University

Biological evidence indicates that the brain atrophy can be involved at the onset of neuropathological pathways of Alzheimer’s disease. However, there is lack of formal statistical methods to perform genetic dissection of brain activation phenotypes such as shape and intensity. To this end, we propose a Bayesian hierarchical model which consists of two levels of hierarchy. At level 1, we develop a Bayesian nonparametric level set model for studying the brain activation region shape. At level 2, we construct a regression model to select genetic variants that are strongly associated with the brain activation intensity, where a spike-and-slab prior and a Gaussian prior are chosen for feature selection. We develop efficient posterior computation algorithms based on the Markov chain Monte Carlo (MCMC) method. We demonstrate the advantages of the proposed method via extensive simulation studies and analyses of imaging genetics data in the Alzheimer’s disease neuroimaging initiative (ADNI) study.

Relationships Between Brain Structural Connectome and Traits
Zhengwu Zhang
University of Rochester

Advanced brain imaging techniques make it possible to measure individuals’ structural connectomes in large cohort studies non-invasively. However, due to limitations in image resolution and pre-processing, questions remain about whether reconstructed connectomes are measured accurately enough to detect relationships with human traits and behaviors. Using a state-of-the-art structural connectome processing pipeline and a few novel dimensionality reduction techniques applied to data from the Human Connectome Project (HCP), we show strong relationships between connectome structure and various human traits. We find that brain connectomes are associated with many traits. Specifically, fluid intelligence, language comprehension, and motor skills are associated with increased cortical-cortical brain connectivity, while the use of alcohol, tobacco, and marijuana are associated with decreased cortical-cortical connectivity.

Session 52: Integrative analysis of multi-view data with applications to precision medicine

Joint Modeling of Multi-System Wearable Data
Vadim Zipunnikov
Johns Hopkins Bloomberg School of Public Health

Physical activity and sleep trackers as well as heart rate monitors are now extensively used to track quality of sleep, levels of physical activity, and disruptions in circadian rhythms in many clinical studies, for example, to reliably monitor and assess post-intervention changes in patient’s status. Thus, mobile technologies has provided an unprecedented opportunity to obtain objective simultaneous assessment of multiple physiological systems in real-time, over weeks or months. We developed a computationally efficient data analytical framework that parsimoniously models multiple biomarkers nested within multiple systems, accounts for interactions of biomarkers within and between physiological systems, and decomposes within-subject and between-subject variability. We illustrate our approach using 7-day of Actiwatch and Actiheart data from 106 participants of Baltimore Longitudinal Study of Aging.
Integrating large-scale sequencing data for cancer classification
Ronglai Shen
Memorial Sloan Kettering Cancer Center
ronglais@gmail.com

In this talk, I will present a pan-cancer analysis of multiple omic data platforms from the Cancer Genome Atlas. Collectively, the multiple types of genomic, epigenomic, and transcriptomic alterations can provide refined subtype classifications and increased power and precision to determine the molecular basis of clinical phenotypes. We identified shared molecular alterations in different cancer types which may indicate related disease etiology and provide unique opportunities to compare treatments and outcome across cancer types. I will also discuss the prognostic relevance of tumor sequencing in a clinical setting. We performed a mutational analysis of > 300 cancer-associated genes in 1,054 patients with metastatic lung adenocarcinoma sequenced prospectively. An integrated prognostic scoring model was developed based on mutational profiling to stratify patients into different risk groups. The integrated prognostic scoring system can facilitate the incorporation of mutational data as a stratification factor in both prospective clinical trials and retrospective/real-world data collections to more precisely describe patient populations, allowing better generalization of the results from such research efforts.

Integrative clustering of multidimensional omics data

• Mengyun Wu¹, Sebastian Teran Hidalgo, Yang Li² and Shuangge Ma

1Yale University
2Renmin University of China
mengyun.wu@yale.edu

Clustering analysis has been a critical component of genetic study on complex diseases. For example, analyzing gene expression data can reveal the interconnections among genes, and clustering samples can serve as the basis of disease subtype identification, risk stratification, as well as many other purposes. With only small sample size of genetic profiling data and the noisy nature of omics data, clustering analysis on a single dataset often leads to unsatisfactory results. In recent profiling studies, a prominent trend is to collect data on multidimensional omics data, including gene expressions and their regulators (copy number alteration, microRNA, methylation, etc.) on the same subjects. In our study, we develop integrative clustering methods for multidimensional omics data based on the regularized estimation and NCut techniques. Extensive simulations show that our proposed methods have significantly improved clustering accuracy. In the analysis of cancer genomic data, biologically sensible findings that are different from the alternatives are made.

An Integrative Graphical Modeling Approach for Multiple Heterogeneous Omics Data

• Anantiza Mcgillivray¹ and George Michailidis²

¹University of Florida Informatics Institute
²University of Florida
amcgillivray@ufl.edu

Rapid development of high-throughput technologies has led to the collection of data across multiple molecular compartments (genomic, transcriptomic, proteomic, etc.) across many different biological conditions. Standard graphical modeling methods are limited in their ability to deal with such heterogeneous data, which may also differ in size and noise levels, emphasizing the need for integrative approaches. In this talk, we present a statistical framework for performing two types of integrative graphical modeling. We first present an integrated approach that jointly estimates multiple Gaussian graphical models from a variety of omics sources. We next extend this approach to allow for both multiple omics sources and varying biological conditions. We discuss the computational procedure for solving the resulting optimization problems, establish theoretical properties of the proposed estimators, and illustrate their performance with an application to TCGA breast cancer data.

Session 53: Real world evidence in the US health care system
Successfully Integrate RWE in Clinical Processes in the Era of Big Data
Kelly Zou
Pfizer Inc.
kelly.zou@pfizer.com

Given the desire to enhance the effectiveness and efficiency of health care systems, it is important to understand and evaluate the risk factors for disease progression, treatment patterns such as medication uses, and utilization such as hospitalization. Statistical analyses via observational studies and data mining may help evaluate patients’ diagnostic and prognostic outcomes, as well as inform policies to improve patient outcomes and to control costs. In the era of big data, real-world longitudinal patient-level databases containing the insurance claims of commercially insured adults, electronic health records, or cross-sectional surveys, provide useful insights to such analyses. Within the healthcare industry, executing rapid queries to inform development and commercialization strategies, as well as pre-specified non-interventional observation studies, are commonly performed. In addition, pragmatic studies are increasingly being conducted to examine health-related outcomes. In this presentation, selective published examples on real-world data analyses are illustrated. Results typically suggest that paying attention to patient comorbidities and pre-index or at index health care service utilization may help identify patients at higher risk and unmet needs for treatments. Finally, fruitful collaborative opportunities exist across different sectors among academia, industry and the government.

Statistical challenges and method-exploration for intermittent missing data occurring concurrently with right censoring:
Quantitative illustration from a real-world study evaluating treatment patterns and unmet needs in adult patients with moderate-to-severe Atopic Dermatitis (AD)

• Zhen Chen¹, Toshio Kimura, Weiming Du¹, Usha Mallya, Abhijit Gadkari and Raymond Miao
1Regeneron Pharmaceuticals, Inc.
need2learn421@gmail.com

The goal of this work is to present a quantitative illustration of deriving an optimal solution for the problem of intermittent missing data occurring concurrently with right censoring in a real-world study. This will be illustrated through data collected from a study evaluating moderate-to-severe Atopic Dermatitis (AD) adult patients’ adeQuacy of Existing Systemic Treatments (AD-QUEST). This AD-QUEST study was designed to evaluate flares (defined as increased itching/redness or new/spreading lesions) in a real-world setting among adults with moderate-to-severe AD treated with systemic agents prior to recent approval of biologics for this population. Adults (≥18 year) with an AD diagnosis from June 2011 to May 2016, and with a prescription for an immunosuppressant, systemic corticosteroid, or phototherapy in the last 6 months, were identified from the Optum Research Database. Moderate-to-severe AD patients (self-assessed based on Rajka-Langeland criteria) took

2018 ICSA Applied Statistics Symposium, New Brunswick, NJ, June 14-17
97
part in a baseline paper survey and monthly web-based surveys over 12 months. There were 801 patients who completed the baseline survey (mean ± SD age: 45±14 years; female: 72%; white: 84%; employed: 79%; AD severity: moderate 74%, severe 26%). At baseline, in the past month, 23.3% of patients reported having 1 flare, 19.7% had 2 flares, and 38.3% had > 2 flares. Of those reporting a flare, 49.6% reported that each flare lasted between 1-3 weeks; 28.0% reported that each flare lasted for ≥3 weeks. 91.6% reported worrying about a flare during the past month. For the monthly surveys, N=473, 455, 482, 466, 476, 451 out of 801 baseline patients completed the assessment at months 1 to 6, respectively. This intermittent missing data type presents a statistical challenge, especially when it occurs concurrently with right censoring. This challenge was identified during protocol development, and simulations were conducted to evaluate different analysis methods such as non-parametric Kaplan-Meier estimator and parametric exposure adjusted method for time to first event estimation, with consideration of missing at random and missing not at random (where patients’ willingness to deliver survey assessment is influenced by their disease conditions).

The statistical simulation results illustrate the performance (as evaluated by mean squared error) of different statistical methods under various assumptions and support the parametric exposure adjusted method as being the optimal among all competing methods; therefore, it was selected as the protocol-specified method for the study.

Illuminating variation in implantable cardiac device use and outcomes with billing claims data

Laura Hatfield1, Sharon-Lise Normand2, Daniel Kramer2 and Rita Volya1
1Harvard Medical School
2Beth Israel Deaconess Medical Center

Cardiac resynchronization therapy pacemakers without (CRT-P) and with (CRT-D) defibrillator backup are used to treat congestive heart failure. Across providers, states, and countries, the choice of which device to implant varies widely, and use of CRT-D devices is much more common in the United States than in other countries. Using fee-for-service Medicare billing claims, we study real-world outcomes of more than 70,000 patients implanted with CRT-D and CRT-P devices. We quantify variation in the tendency to implant one device versus the other at the facility level and we exploit this variation to compare outcomes of patients receiving the two devices. We compare the mortality, hospitalization, and re-operation outcomes of patients receiving each using both regression and instrumental variables (IV) analyses. When aggregated to the state level, CRT-P implant probability is not consistently related to other measures of health care utilization intensity. In our IV analysis, we find that patients who received CRT-P devices had no worse survival and lower probability of admissions and re-operations in the year following implant compared to patients who received CRT-D implants.

Real world challenges in the design of pragmatic trials

Valentina Zubek
Boehringer Ingelheim Pharmaceuticals, Inc.
valentina.zubek@boehringer-ingelheim.com

Pragmatic randomized clinical trials (PrCTs) extend the framework of clinical trials to the real world clinical practice setting. Some of the characteristics that set PrCTs apart are: minimal inclusion/exclusion criteria; fewer scheduled visits mimicking the real-world interaction of patients and physicians (as opposed to the regular schedule of clinical visits); a variety of data sources such as clinician-collected data, electronic health records and insurance claims; and outcomes that extend beyond efficacy and safety to include health care resource utilization and costs, and that are relevant to patients, physicians and clinical decision makers. These characteristics may be interpreted as both strengths and weaknesses for the PrCTs. For example, the limited eligibility criteria would simplify the enrollment process, lead to a heterogeneous, real-world population which helps increase the generalizability of results, but this population may also exhibit smaller treatment effects and higher variability. Likewise, while fewer scheduled visits reflect real-world interactions, they may impact adherence, which will affect intention-to-treat analysis vs. on-treatment analysis. One of the most important features of PrCTs is randomization (often, cluster randomization), which ensures a solid statistical foundation for comparing interventions. PCTs are often open label, which brings more biases. Most importantly, PrCTs are designed to provide real-world evidence about healthcare interventions while preserving randomization, therefore guiding the decision making process of patients, clinicians, payers and health policy decision makers.

Session 54: Statistical Inference for Discrete and Categorical Data

Evaluating vaccine efficacy variation by type of infecting pathogen using Dempster-Shafer (DS) Multinomial models

Paul Edlefsen1, Robin Gong2 and Jan Hannig3
1Fred Hutchinson Cancer Research Center
2Harvard University
3University of North Carolina at Chapel Hill
pedlefse@fredhutch.org

Evaluating vaccine efficacy as it varies across the features ("marks") of an infecting pathogen is called "sieve analysis" and it has an important role in vaccine design and evaluation. For example the viruses that cause dengue fever comprise four types of DENV pathogen, and the WHO call for a vaccine against dengue fever explicitly requires equal and strong protection against all four types. A sieve analysis of the new Sanofi dengue fever vaccine found that among younger recipients the vaccine showed differential and lower efficacy (a "sieve effect") against one of these four types, and as a result the vaccine has been licensed (in Brazil and elsewhere) with an exclusion for younger recipients. We have previously presented statistical methodology to evaluate sieve effects based on dichotomous endpoints using Dempster-Shafer binomial models, and here we present the extension to general multinomial models such as are required to evaluate efficacy against the four DENV types. We employ a new efficient implementation of the Dempster 1972 "simplex method" as well as a new implementation of an unpublished "Poisson box projection" method for DS multinomial inference. This is joint work with Ruobin Gong and Jan Hannig.

Change Point Detection for Poisson Time Series Images with Applications to Astronomy and Astrophysics

Thomas Lee
University of California, Davis
tcmlee@ucdavis.edu

In this talk we report our on-going work on detecting change points in a time series sequence of Poisson images. We develop fast algorithms and establish consistency results. We apply our methodology
to a time lapse movie of the Chandra observations of the Orion Nebula cluster.
This is joint work with Vinay Kashyap and Cong Xu.

Poisson is enticing, but this preprocessed long-tail fish has fewer bones...
Xiao-Li Meng
Harvard University

The standard Poisson model is enticing: a single, easily estimable mean parameter governs the entire Poisson distribution. The model’s validity can be established mathematically for many physical observations, such as photon counts from astronomical sources. However, Poisson’s elegance also limits its applicability when the counts are preprocessed, or when the mean parameter in a Poisson regression cannot be modeled accurately. A log-normal model offers a separate, distinct variance parameter to accommodate measurement errors and model imperfection. To maintain the mean model on the original scale, a half-variance correction is typically needed because expectation is not invariant to the log transform. An intriguing byproduct of the half-variance correction is a nonlinear, self-shrinkage phenomenon in variance estimation. Further extension via log-t models provides an effective strategy for dealing with outliers. This talk illustrates all these points using the problem and data from "Calibration Concordance for Astronomical Instruments via Multiplicative Shrinkage" (JASA, to appear), a collaboration between astrophysicists (Herman Marshall of MIT and Vinay Kashyap of Harvard) and statisticians (Yang Chen, Xiao-Li Meng, Xufei Wang, and David van Dyk).

Session 55: Modern Statistical Development for Biomedical Big Data

Bayesian Large Scale Inference for Time to First Event of Multivariate Ordinal Survival Outcomes with Application to TennCare Cohort Study
Qingxia Chen1, David Schlueter2, Christopher Fonnesbeck1 and Pingheng Wu1
1Vanderbilt University Medical Center
2Vanderbilt University

Motivated by a large size TennCare Cohort Study, we develop a scalable Bayesian framework to accommodate time to first event of multivariate survival outcomes with ordinal severity and model the multivariate survival outcomes using a flexible gamma frailty transformation model. Using an additional data source correlating the multivariate survival outcomes with ordinal severity scores, we provide a systematic and flexible way to determine the overall direction of the effect size over multivariate survival events. A computationally efficient algorithm based on variational inference is used to scale the Bayesian inferential scheme to big data. Bayesian model selection procedures are further developed to determine the most proper and pragmatic transformation. Numerical simulations are conducted to evaluate the validity of the method and variational algorithm. The proposed method is further applied to a large cohort from the Tennessee Asthma Bronchiolitis Study.

Online Updating of Survival Analysis in the Big Data Setting
Jing Wu1, Elizabeth Schifano, Jun Yan2, Yishu Xue and Ming-Hui Chen2
1University of Rhode Island
2University of Connecticut

When large amounts of survival data arrive in streams, conventional estimation methods may become infeasible since they require storage of all the risk sets at each accumulation point. In this paper, we develop online updating methods for carrying out survival analysis under the Cox proportional hazards models. Specifically, we propose online-updating estimators as well as their corresponding standard errors for both the regression coefficients and the baseline hazard function. Theoretical properties of the estimators are examined in detail. An extensive simulation study is conducted to examine empirical performance of the proposed estimators. A large colon cancer data set from SEER is analyzed to further demonstrate the proposed methodology.

Estimating Causal Log-Odds Ratio Using Case-Control Sample and Its Application in the Pharmaco-Epidemiology Study
Donglin Zeng
University of North Carolina at Chapel Hill

One of the most important goals in pharmaco-epidemiology studies is to understand the causal relationship between drug exposures and their clinical outcomes, including drug-induced adverse events (ADEs). However, this goal needs to resolve the following difficulties: (1) most of pharmaco-epidemiology data are observational and confounding is largely present due to many comedications; (2) data are often collected from a big database using a matched case-control design; (3) data analysis needs to handle a big dataset size which cannot be handled using existing packages. In this paper, we tackle these challenges both methodologically and computationally. First, we propose a conditional causal log-odds ratio definition to characterize individual causal effects of drug exposures on binary ADE status. We then provide sufficient conditions with control-based propensity scores to estimate this causal log-odds ratio using a case-control design. Computationally, we implement dimension reduction to avoid fitting a model with high dimensional confounders and adopt meta-analytic procedures to analyze big data sets. We conduct extensive simulation studies to demonstrate superior performance to existing methods. Finally, we apply the proposed method to analyze one myopathy data ADEs set from The Indiana Network for Patient Care which has over 6 million records, and study the causal relationship between each of 210 drugs and myopathy ADEs.

Quantile Decision Trees and Forest with its application for predicting the risk (Post-Traumatic Stress Disorder) PTSD after experienced an acute coronary syndrome
Ying Wei1, Huichen Zhu1 and Ian Kronish1
1Columbia University

Classification and regression trees (CART) are a classic statistical learning method that efficiently partitions the sample space into mutually exclusive subspaces with the distinctive means of an outcome of interest. It is a powerful tool for efficient subgroup analysis and allows for complex associations and interactions to achieve high prediction accuracy and stability. Hence, they are appealing tools for precision health applications that deal with large amounts of data from EMRs, genomics, and mobile data and aim to provide a transparent decision mechanism. Although there is a vast literature on decision trees and random forests, most algorithms identify subspaces with distinctive outcome means. The most vulnerable or high-risk groups for certain diseases are often patients with extremely high (or low) biomarker and phenotype values. However, means-based partitioning may not be effective for identifying pa-
patients with extreme phenotype values. We propose a new regression tree framework based on quantile regression (Koenker and Bassett, 1978) that partitions the sample space and predicts the outcome of interest based on conditional quantiles of the outcome variable. We implemented and evaluated the performance of the conditional quantile trees/forests to predict the risk of developing PTSD after experiencing an acute coronary syndrome (ACS), using an observational cohort data from the REACTIONS to Acute Care and Hospitalization (REACH) study at New York Presbyterian Hospital. The results show that the conditional quantile based trees/forest have better discrimination power to identify patients with severe PTSD symptoms, in comparison to the classical mean based CART.

Session 56: New advanced methods in functional data analysis

Optimal weighting schemes for longitudinal and functional data
\*Xiaoke Zhang\^1 and Jane-Ling Wang\^2
\^1George Washington University
\^2University of California, Davis
\texttt{xkzhang@gwu.edu}

We propose optimal weighting schemes for both mean and covariance estimations for functional data based on local linear smoothing such that the L2 rate of convergence is minimized. These schemes can self-adjust to the sampling plan and lead to practical improvements.

A simultaneous confidence band for variance function based on dense functional data
Suojin Wang
Texas A & M University
\texttt{sjwang@stat.tamu.edu}

A two-step reconstruction-based moment estimator and an asymptotically correct smooth simultaneous confidence band are proposed for the heteroscedastic variance function of dense functional data. Step one involves spline smoothing for individual trajectory reconstructions and step two employs kernel regression on the individual squared residuals to estimate each trajectory variability. A moment estimator is then constructed for the variance function by using these estimated trajectory variabilities. The variance estimator is shown to be oracally efficient in the sense that all individual trajectories are estimated as efficiently as if they were known by ‘oracle’, with which an asymptotically correct SCB is established for the variance function. Our simulation results support the asymptotic theory. As an illustration, the proposed variance function estimator and its SCB are applied to analyze two real data sets.

Supervised Principal Component Regression for Functional Data with High Dimensional Predictors
\*Dehan Kong\^1, Xinyi Zhang\^1 and Qiang Sun\^1
\^1University of Toronto
\texttt{kongdehan@utstat.toronto.edu}

Motivated by functional magnetic resonance imaging studies, we study the effect of clinical/demographic variables on the dynamic functional structures, which plays a key role in understanding brain functionality. To this end, we propose the supervised principal component regression for functional data with possibly high dimensional clinical variables. Compared with its classical counterpart, the principal component regression, the proposed methodology relies on a newly proposed integrated residual sum of squares for functional data and makes use of the association information directly. It can be formulated as a sequence of nonconvex generalized Rayleigh quotient optimization problems, which turn out to be NP-hard and thus computational intractable. Utilizing the invariance property of linear subspaces under rotations, we then reformulate the problem into a simultaneous sparse linear regression problem. Formally, we show that the reformulated problem can recover the same subspace as if the original sequence of nonconvex problems were solved. Statistically, the rate of convergence for the obtained estimators is established. Numerical studies and an application to the Human Connectome Project lend further support to the proposed methodology.

Session 57: Dimension reduction, variable selection and their applications

Sufficient Variable Selection Using Independence Measures for Continuous Responses
\*Xiangrong Yin\^1, Baoying Yang and Nan Zhang
\^1University of Kentucky
\texttt{yinxiangrong@uky.edu}

In this talk, we propose two sufficient variable selection procedures: one-stage and two-stage approaches using independence measures for continuous responses. Although any independence measure can be used, we use distance correlation and Hilbert-Schmidt IndependenceCriterion correlation to illustrate the two procedures. By comparing with some existing marginal screening methods, we show the advantages of the proposed procedures through the simulation studies and a real data analysis. Our procedures are model-free and robust against model mis-specification, especially useful, when the active predictors are marginally independent of the response.

Some Dimension Reduction Strategies for the Analysis of Survey Data
\*Derek Young and Jiaying Weng
University of Kentucky
\texttt{derek.young@uky.edu}

In the era of big data, researchers interested in developing statistical models are challenged with how to achieve parsimony. Sufficient dimension reduction has emerged as an area of broad and current interest. These types of dimension reduction strategies have been
Spatial envelope models

Qiwen Wang
Virginia Commonwealth University
qwang3@vcu.edu

Dimension reduction provides a useful tool for analyzing high dimensional data. The recently developed Envelope method is a parsimonious version of the classical multivariate regression model through identifying a minimal reducing subspace of the responses. However, existing envelope methods assume an independent error structure in the model. While the assumption of independence is convenient, it does not address the additional complications associated with spatial or temporal correlations in the data. In this talk, we introduce a Spatial Envelope model for dimension reduction in the presence of dependencies across space. We study the asymptotic properties of the proposed estimators and show that the asymptotic variance of the estimated regression coefficients under the spatial envelope model is smaller than that from the traditional maximum likelihood estimation. Furthermore, we present a computationally efficient approach for inference. The efficacy of the new approach is investigated through simulation studies and an analysis of an Air Quality Standard dataset from the Environmental Protection Agency (EPA).

Efficient integration of sufficient dimension reduction and prediction in discriminant analysis

Xin Zhang
Florida State University
henry@stat.fsu.edu

Sufficient dimension reduction (SDR) methods are popular model-free tools for pre-processing and data visualization in regression problems where the number of variables is large. In discriminant analysis, there is usually no guarantee of improvement in classification accuracy by a reduce-and-classify approach, mainly due to the different natures of the two stages. Envelope methods, on the other hand, construct targeted dimension reduction subspaces that often improve efficiency in estimating a multivariate parameter. However, little is known about how to construct envelopes in discriminant analysis models. In this paper we introduce the notion of the Envelope Discriminant Subspace (ENDS) as a natural inferential and estimative object in discriminant analysis that incorporates these considerations. We develop the ENDS estimators that simultaneously achieve sufficient dimension reduction and classification. Consistency and asymptotic normality of the ENDS estimators are established, where we carefully examine the asymptotic efficiency gain under the classical linear and quadratic discriminant analysis models. Simulations and real data examples show superb performance of the proposed method.

Session 58: Analysis and computation for complex models

Bayesian functional quantile regression

Yusha Liu1, Jeffrey Morris2 and Meng Li3

1 Rice University
2 University of Texas M.D. Anderson Cancer Center
3 University of California, San Diego

We propose a unified Bayesian functional quantile regression framework which simultaneously performs quantile regression by adopting an asymmetric Laplace working likelihood and adaptively regularizes the functional regression coefficients by projecting them into a basis space and employing an appropriate shrinkage prior on the basis coefficients. Our framework is highly general in that any basis functions and shrinkage priors can be chosen, depending on the characteristics of the functional data to be analyzed. We developed an efficient Gibbs sampler to fit this fully Bayesian hierarchical model automatically with no tuning required, which yields posterior samples that can be used to perform pointwise or joint inference on any model quantity of interest. Our approach is computationally efficient and can handle functional data observed on grids of hundreds to thousands. Simulation studies demonstrate that our proposed model achieves substantially improved performance than competing methods, especially for more extreme quantiles. We have applied our approach to a mass spectrometry proteomics dataset.

Validation of Approximate Bayesian Computation on Posterior Convergence

Wen Tao Li1 and Paul Fearnhead2
1 Newcastle University
2 Lancaster University
wenli@newcastle.ac.uk

The likelihood function is key to many statistical methods, including maximum likelihood estimation and Markov chain Monte Carlo. But for many realistic models in modern applications, it is difficult to evaluate the likelihood functions due to high model complexity, e.g., models with nonlinear dynamics and latent structure. Approximate Bayesian computation (ABC) method implements Bayesian inference without evaluating the likelihood function, only requiring the ability to simulate pseudo datasets from the model. More specifically, the inference is based on simulated datasets that are close to the observed data. ABC has been popular within population genetics and ecology over a decade, and has recently found wide applications in other areas involving financial time series and stochastic differential equation. This talk will discuss the limitation of standard implementation of ABC, in the sense that it can perform well in terms of point estimation, but will over-estimate the uncertainty about the parameters. If we use the regression correction of Beaumont et al. (2002), then ABC can also accurately quantify this uncertainty, hence achieving the same asymptotic accuracy as the likelihood-based methods.

Learning Temporal Evolution of Spatial Dependence

Shiwei Lan
Department of Computing + Mathematical Sciences, California Institute of Technology
lanzithinking@gmail.com

Spatio-Temporal data are ubiquitous in science and engineering applications. They feature challenges including (i) big (high volume) data (e.g. medical records), (ii) high dimensions (of the state space) (e.g. climate data) and (iii) complex structures (e.g. network data). Learning temporal evolution of spatial dependence in these data is important to uncover the dynamic relationship among multiple time series from different locations. This could have impact in informing the brain disease mechanism, genome-wide association in evolution, and regional climate change etc.

In this talk, I will present a novel Bayesian framework to model the dynamic correlation (covariance) among multiple processes. The
method is based on modeling the correlations as products of vectors on unit spheres. This geometric approach not only enables us to induce flexible prior distributions for correlation (covariance) matrices, but it also provides us an efficient inference algorithm called Spherical Hamiltonian Monte Carlo to handle the norm constraints. Unit-vector Gaussian process priors are then introduced in order to capture the evolution of correlation among multiple time series. I will use a simulated periodic process to illustrate the validity and effectiveness of the proposed framework, and analyze the local field potential signals (record from the hippocampus of a rat in a sequence memory experiment) using the model to study the dynamic brain connectivity during such cognitive process. In the end I will highlight full nonparametric generalizations (ongoing) of the proposed model in both time and space domains.

submitted paper: https://arxiv.org/abs/1711.02869

demo of the dynamic brain connectivity study: https://www.youtube.com/watch?v=NMUUic0IdmM

Session 59: Hierarchical generalized linear models in practice

Optimality of the maximum h-likelihood estimators
Youngjo Lee
Seoul National University
youngjo@snu.ac.kr

The h-likelihood has been proposed for statistical inference on random-effect models. In this presentation, the properties of the mode of the h-likelihood for inference on random effects of clustered data will be introduced. To define optimality in random-effect predictions, several foundational concepts of statistics such as likelihood, unbiasedness, consistency, confidence distribution and the Cramer-Rao lower bound are extended. Exact probability statements about interval estimators for random effects can be made asymptotically without a prior assumption. With large samples, the optimization procedure based on h-likelihood provides an efficient way of computing relevant quantities for the Bayesian inference without having to resort to MCMC samples. We also discuss when random-effect prediction is beneficial for fixed effect estimation.

Keywords: Maximum likelihood estimator; Best linear unbiased predictor; Bayes estimator; Bartlett identities; Cramer-Rao bound; Confidence distribution

H-likelihood Inference for Survival Models with Random Effects
Il do Ha
Pukyong National University
idhal1353@pknu.ac.kr

In this talk, we present recent works on extensions of the hierarchical (or h-) likelihood (Lee and Nelder, 1996) to time-to-event data (survival data), which overcomes various challenges due to incomplete observations caused by censoring, truncation and competing events, and also present further extensions of existing works, such as complicated structured frailties (or random effects) and joint models for repeated measures and time-to-events. In particular, we show that the h-likelihood approach gives a useful methodology for interval estimation of the individual frailty and variable selection of covariates in survival models (e.g. proportional hazards and AFT models) with random effects. We also demonstrate via the h-likelihood how to make inference various random-effect survival models using time-to-event data from multi-center clinical trials, and the inferential results are also compared with those of copula survival models which are useful for modeling clustered survival data.

Power and false discovery rates of HGLMs and other statistical models analyzing RNAseq data

Dongseok Choi1, Maengseok Noh2, Jiwoong Kim3 and Youngjo Lee3
1Oregon Health & Science University
2Pukyong National University
3Seoul National University
choi@ohsu.edu

RNAseq has become the standard technology in gene expression studies in the past few years. It is considered superior to microarrays that used to be the choice of technology in the 2000s. There are still many challenges in analyzing RNAseq data and its analytic pipelines require multiple steps and combine the methods of bioinformatics and biostatistics. Since RNAseq data are typically processed to counts for downstream statistical analyses, there have been active developments of statistical models based on negative binomial regression for high dimensional counting data. The main statistical challenge is estimating complex dispersion structure in high dimensional counting data. Some popular choices are implemented in R packages such as DESeq2 and edgeR. First introduced by Lee and Nelder in 1996, the hierarchical generalized linear models (HGLMs) allow random effects in linear predictors for mean and dispersion and can be efficient in estimating complex dispersion structure of RNAseq data. In this presentation, we will review a brief history of advancement of statistical methods for RNAseq data and compare their power and false discovery rates by simulations.

Data analysis using MDHGLM package in R

Maengseok Noh
Pukyong National University
msnoh@pknu.ac.kr

We present the mdhglm R package for fitting multivariate double hierarchical generalized linear models (DHGLMs) using the h-likelihood method. This package allows different models for multivariate response variables where each response is assumed to follow DHGLMs. With DHGLMs, the mean and dispersion parameters for the variance of random effects and residual variance (overdispersion) can be modeled as random-effect models. Furthermore, the mdhglm package allows various correlation structures among the random effects and saturated random effects in DHGLMs of different multivariate response variables and structural equation models. With several data-sets, we illustrate the use of the mdhglm package.

Session 60: Building Successful Sponsor/CRO Relationship - A Biostatistical Perspective

Statistical Considerations: Best practices for Sponsor/CRO Relationships

Patrick Larson
Merck & Co., Inc.
patrick.larson@merck.com

Over the last decade sponsors are using CROs to perform statistical analysis in the conduct of clinical trials that were traditionally kept in-house. Outsourcing of this information, depending on the relationship, can take on many forms. For lack of a better term this could include functional outsourcing which might include only the statistical analysis/programming of the data from ongoing trials or completed studies in preparation of the clinical study reports (CSR).
Additionally this relationship between the sponsor and CRO could also be expanded to the protocol writing, site management, statistical programming and analysis as well as the finalization of the clinical study report. For the latter scenario, in this talk this will be referred to as a fully outsourced trial. In this presentation, I will provide some best practices and considerations for each of these types of relationships based on some personal experiences in the setting of the design, analysis, reporting of clinical pharmacology trials.

**Key Components for a Successful Sponsor/CRO Strategic Partnership**

**Jianrong Li**

PPD

[jianrong.li@ppdi.com](mailto:jianrong.li@ppdi.com)

It is imperative that the cost of Research and Development of new and innovative therapies be optimized to bring greater health improvements to patients worldwide. To meet this challenge, both sponsors and CROs are seeking solutions to advance products more efficiently. In addition to technology and innovations, sponsors and CROs are also looking to create deeper and more strategic partnerships to integrate resources for more streamlined processes and greater efficiencies. Developing strategic partnership allows sponsors to take their CRO partners into their holistic R&D planning. Strategic partnerships also provide CROs with the visibility into their sponsors’ and partners’ R&D pipelines for the ability to proactively plan workforces. Under robust and engaging governance structures and comprehensive operational models, the strategic sponsor/CRO partnership will eliminate overlapping activities and produce efficient drug development processes. In this presentation, we will discuss the key components for ensuring successful and sustainable strategic partnerships using a real world example with the focus on Bio-statistics and Programming services.

**Functional Service Outsourcing Trends and Experience**

**Tiepu Liu**

FMD K&L, Inc.

tiepu.liu@klserv.com

Global pharmaceutical R&D outsourcing to CROs is still trending up, with an estimated value at USD 40 billion in 2018. Sponsors outsource to CROs for different services in various forms over time. Functional Service Provider (FSP) model is commonly used, especially for data management and statistical analysis services. FSP model is practiced by full service CROs as well as specialty service CROs. Large and small sponsor companies apply this model for data and analysis services. Usually CRO manages employment of the resource and dedicates that to the sponsor at FTE basis. Sponsors can directly manage the project work for these FTEs without worrying about internal head count or ramping up or drawing down hassle. Broadly speaking sponsors also use FSP model combining dedicated resource and flexible resource and applying both deliverable based pricing and time and material based costing. Such a sponsor and FSP partnership allows more flexibility and close collaboration to improve quality and efficiency. In this presentation we will share experience and thoughts from real world FSP based data and analysis services.

**Session 61: Statistical and Regulatory Issues in Human Drug Abuse Studies**

**Statistical Issues in Human Drug Abuse Studies**

**Qianyu Dang**

U.S. Food and Drug Administration

According to National Center for Health Statistics, there were more than 64,000 drug overdose deaths in 2016, which makes it the leading cause of accidental death in the US. Among them, prescription opioid painkillers caused 14,400 overdose deaths. Human drug abuse study is an important component in evaluating whether or the magnitude of a drug produces subjective responses related to abuse potential through the drug effects on the human central nerve system (CNS). Currently there are mainly two types of pharmacodynamic (PD) clinical studies: The human abuse potential study (HAP) and human abuse study for abuse-deterrent formulated opioids (ADF).

While HAP studies for ADF products are considered as efficacy studies for their abuse-deterrent properties, general HAP studies are safety studies. For both types of studies, primary hypothesis testing involves comparison between different treatments for outcomes like drug liking and take drug again. Mixed model is the first choice of testing, but the model assumptions such as homoscedasticity and normality of residuals have to be checked before applying the model. If the model assumptions are violated, the testing of hypothesis has been carried through paired t-test or non-parametric approaches instead. But those approaches also have their own assumptions. Further more, for non-parametric test, the power is much lower and the hypothesis itself was different. In this research, we evaluated the impact of violations on different assumptions in the mixed model and their impact on the testing results through numerical simulations based on real data from HAP studies.

**Statistical and Regulatory Issues in Human Drug Abuse Studies**

*Catherine Mills and Denise Milovan*

Syneos Health

catherine.mills@syneoshealth.com

catherine.mills@syneoshealth.com

Following the 2010 release of a "Draft Guidance: Assessment of Abuse Potential of Drugs", the FDA issued a final guidance in January 2017.2 The draft guidance provided recommendations for the assessment of the abuse potential of novel central nervous system acting drugs with the final guidance including clarifications concerning the strategic decisions associated with the design of appropriate studies for the investigation of abuse potential. A separate FDA guidance specific to the investigation of abuse deterrent opioids was released in 2015.3 Both the 2015 and 2017 guidelines provide distinct recommendations for the statistical evaluation of human abuse potential (HAP), which require careful consideration. Methodological considerations for future clinical studies may need to consider both guidances for the statistical evaluation of new chemical entities with analgesic properties, as well as for potential product labeling recommendations related to abuse deterrence. Statistical approaches outlined in the 2017 guidance will be discussed in light of Syneos Health’s standardized methods for data analysis developed through the company’s extensive experience with the assessment of subjective responses to novel and known drugs of abuse, as well as the implementation of guidance documents, and direct feedback from FDA reviewers. The methods for analyzing HAP studies, including testing of residuals for normality, will be presented. Traditionally, the determination of the abuse potential of a new molecular entity included the evaluation of the primary Drug Liking visual analog scale (VAS) Emax endpoint on the following hypotheses: 1. Positive control drug versus placebo: H0: $\mu_C - \mu_P \leq 0$ vs. Ha: $\mu_C - \mu_P > 0$. 2. Test drug versus positive control drug: H0: $\mu_T - \mu_C \geq 0$ vs. Ha: $\mu_T - \mu_C < 0$. 3. Test drug versus placebo: H0: $\mu_T - \mu_P = 0$ vs. Ha: $\mu_T - \mu_P \neq 0$. The 2017 guidance revises the requirements for sample size, statistical model, qualitative vs. quantitative correlations, and the selection of safety margins for hyp-
Abstracts


Overview of Human Abuse Potential Studies as Applied to Abuse Deterrent Assessment
James Tolliver
U.S. Food and Drug Administration
james.tolliver@fda.hhs.gov

The abuse of pharmaceutical opioid products constitutes a significant health threat in the United States. Such products may be abused by oral, intranasal (snorting), inhalation (smoking), or intravenous administration. To facilitate abuse, opioid products may be manipulated by cutting, grating, crushing, or grinding. One approach in attempting to reduce opioid abuse is the pharmaceutical development of abuse-deterrent formulations that are difficult to manipulate and thereby deter abuse. Under a New Drug Application (NDA) these products undergo a pre-market abuse-deterrent assessment to determine if such products, compared to non-abuse deterrent formulations, once out on the market will likely be less abused by selected routes of administration. An important component to this assessment is the human abuse potential study in which the effect of manipulation on subjective reinforcing effects is evaluated. Such studies include a Screening Phase, Naloxone Challenge Test, Drug Discrimination Test, Treatment Phase, and Follow-Up Phase. Subjects consist of non-dependent recreational opioid users having experience in the abuse of opioids by the route of administration under study. Treatments include but may not be limited to placebo, active non-abuse deterrent opioid comparator (immediate or extended release), and manipulated abuse deterrent test product. Selected pharmacokinetic parameters of the active opioid are determined from periodic blood sampling. Subjective reinforcing effects are determined using standardized visual analog scales (VAS) for such measures as Drug Liking, High, Take Drug Again, and Overall Drug Liking. In the case of intranasal studies nasal tolerability and percentage of dose insufflated may also be determined. For the different subjective measures validity is determined by the comparison of the positive control versus placebo. The primary comparison is that of the manipulated test drug versus the positive comparator. Statistically significant and numerically relevant reduction in subjective effects is considered to be predictive of a possible abuse-deterrent effect once a given abuse-deterrent product in on the market.

Session 62: Statistical Innovations in Big Data Analysis

Parallel Markov Chain Monte Carlo Methods for Bayesian Analysis of Big Data
Erin Conlon
University of Massachusetts Amherst
 econlon@mathstat.umass.edu

Recently, new parallel Markov chain Monte Carlo (MCMC) methods have been developed for massive data sets that are too large for traditional statistical analysis. These methods partition big data sets into subsets, and implement parallel Bayesian MCMC computation.
independently on the subsets. The posterior MCMC samples from the subsets are then joined to approximate the full data posterior distributions. Current strategies for combining the subset samples include averaging, weighted averaging and kernel smoothing approaches. Here, I will discuss our new method for combining subset MCMC samples that directly products the subset densities. While our method is applicable for both Gaussian and non-Gaussian posteriors, we show in simulation studies that our method outperforms existing methods when the posteriors are non-Gaussian.

More Efficient Estimation for Logistic Regression with Optimal Subsample
Haiying Wang
University of Connecticut
haiying.wang@uconn.edu

Facing large amounts of data, subsampling is a practical technique to extract useful information. For this purpose, Wang et al. (2017, JASA) developed an Optimal Subsampling Method under the A-optimality Criterion (OSMAC) for logistic regression that samples more informative data points with higher probabilities. However, the original OSMAC estimator use inverse of optimal subsampling probabilities as weights in the likelihood function. This reduces contributions of more informative data points and the resultant estimator may lose efficiency. In this paper, we propose a more efficient estimator based on OSMAC subsample without weighting the likelihood function. Both asymptotic results and numerical results show that the new estimator is more efficient. In addition, our focus in this paper is inference for the true parameter, while Wang et al. (2017, JASA) focuses on approximating the full data estimator. We also develop a new algorithm based on Poisson sampling, which does not require to approximate the optimal subsampling probabilities all at once. This is computationally advantageous when available random-access memory is not enough to hold the full data. Interestingly, asymptotic distributions also show that Poisson sampling produces more efficient estimator if the sampling rate, the ratio of the subsample size to the full data sample size, does not converge to 0. We also obtain the unconditional asymptotic distribution for the estimator based on Poisson sampling.

Joint Principal Trend Analysis for Longitudinal High-dimensional Data Integration
Yuping Zhang1 and Zhengqing Ouyang2
1University of Connecticut
2The Jackson Laboratory for Genomic Medicine
yuping.zhang@uconn.edu

Recent advances in genomic medicine resulted in accumulated longitudinal genomic data, where patients were monitored over time, their biological samples were collected at multiple time points, and the corresponding molecular profiles were measured through high-throughput assays. Longitudinal high-dimensional datasets from genomic and biomedical research may have complicated correlation structures or irregular covariance structures. Consequently, such characteristics raise challenges on dimension reduction and feature selection of longitudinal genomic data. We present a new statistical method to integrate multiple sources of information for better knowledge discovery in diverse dynamic biological processes. We demonstrate the utility of our method through simulations and applications to gene expression data of the mammalian cell cycle and longitudinal transcriptional profiling data in response to influenza viral infections.

On adaptive weighting of omics meta-analysis
Yongseok Park

meta-analysis methods have been widely used to combine results from multiple clinical or genomic studies to increase statistical power and ensure robust and accurate conclusion. Adaptively weighted Fisher’s method (AW-Fisher), initially developed for omics applications but applicable for general meta-analysis, is an effective approach to combine p-values from K independent studies and to provide better biological interpretation by characterizing which studies contribute to meta-analysis. AW-Fisher also suffers from lack of variability estimate of AW weights and fast p-value computation. In this paper, we apply bootstrapping to construct a variability index for the AW-Fisher weight estimator and a co-membership matrix to categorize (cluster) differential expressed genes based on their meta-patterns for intuitive biological investigation. We also develop an importance sampling scheme with spline interpolation to increase accuracy and speed of p-value calculation. The superior performance of the proposed methods is shown in simulations as well as two real omics meta-analysis applications to demonstrate insightful biological findings.
include data-dependent priors without damaging the inferential integrity. This data-dependent prior can be viewed as an initial "distribution estimate" of the target parameter which is updated with the results of the ACC method. A general strategy for constructing an appropriate data-dependent prior is also discussed and is shown to often increase the computing speed while maintaining statistical guarantees. We supplement the theory with simulation studies illustrating the benefits of the ACC method, namely the potential for broader applications than ABC and the increased computing speed compared to ABC.

**A Fast Score Test for Generalized Mixture Models**

*Huichen Zhu*, Gen Li, and Eric Lock

1 Columbia University
2 University of Minnesota

In biomedical studies, testing for homogeneity between two groups, where one group is modeled by mixture models, is of great interest. This paper considers the semiparametric exponential family mixture model proposed by Hong et al 2017, and studies the score test for homogeneity under this model. The score test is nonregular in the sense that nuisance parameters disappear under the null hypothesis. To address this difficulty, we propose a modification of the score test, so that the resulting test enjoys the Wilks phenomenon. In finite samples, we show that with fixed nuisance parameters the score test is locally most powerful. In large samples, we establish the asymptotic power functions under two types of local alternative hypotheses. Our simulation studies illustrate that the proposed score test is powerful and computationally fast. We apply the proposed score test to an UK ovarian cancer DNA methylation data for identification of differentially methylated CpG sites.

**Generalized Integrative Principal Component Analysis for Multi-Type Data with Block-Wise Missing Structure**

*Huichen Zhu*, Gen Li, and Eric Lock

1 Columbia University
2 University of Minnesota

High-dimensional multi-source data are encountered in many fields. Despite recent developments on the integrative dimension reduction of such data, most existing methods cannot easily accommodate data of multiple types (e.g., binary or count-valued). Moreover, multi-source data often have block-wise missing structure, i.e., data in one or more sources may be completely unobserved for a sample. The heterogeneous data types and presence of block-wise missing data pose significant challenges to the integration of multi-source data and further statistical analyses. In this paper, we develop a low-rank method, called Generalized Integrative Principal Component Analysis (GIPCA), for the simultaneous dimension reduction and imputation of multi-source block-wise missing data, where different sources may have different data types. We also devise an adapted BIC criterion for rank estimation. Comprehensive simulation studies demonstrate the efficacy of the proposed method in terms of rank estimation, signal recovery, and missing data imputation. We apply GIPCA to a mortality study. We achieve accurate block-wise missing data imputation and identify intriguing latent mortality rate patterns with sociological relevance.

**Session 64: An introduction to BFF: new statistical inference tools for data science**

Could descendants of Jeffreys, Fisher and Neyman become best friends forever?

*Jan Hannig*

University of North Carolina at Chapel Hill

Foundations of statistical inference have been fraught with strong feelings, personal and professional disagreements. However, in the last decade a more collaborative approach has emerged. In this talk I will survey current developments of fiducial inference, inference models, confidence distributions, and objective Bayes, and discuss how we can together address some important questions in statistics and data science.

**Session 65: New Advances in Functional Data Analysis**

Principal Component Analysis for Functional Data on Riemannian Manifolds and Spheres

*Xiongtao Dai* and Hans-Georg Mueller

1 University of California, Davis
2 Texas A&M University
3 University of Minnesota

Functional data analysis on nonlinear manifolds has drawn recent interest. Sphere-valued functional data, which are encountered for example as movement trajectories on the surface of the earth, are an important special case. In this talk, we consider a principal component analysis for smooth Riemannian manifold-valued functional data, which respects the intrinsic geometry of the manifold. Riemannian functional principal component analysis (RFPCA) is carried out by first mapping the manifold-valued data through Riemannian logarithm maps to linear tangent spaces around the time-varying Frechet mean function, and then performing a classical multivariate functional principal component analysis. Representations of the sample functions and the eigenfunctions on the original manifold are then obtained with exponential maps. We derive a central limit theorem for the mean function, as well as root-n uniform convergence rates for other model components. Our applications include a novel framework for the analysis of longitudinal compositional data, achieved by mapping longitudinal compositional data to trajectories on the sphere, illustrated with longitudinal fruit fly behavior patterns. Riemannian functional principal component analysis is shown to be superior in terms of trajectory recovery and predictive power in comparison to an unrestricted method.

Covariance Estimation and Principal Component Analysis for Spatially Dependent Functional Data

*Yehua Li*

University of California, Riverside

We consider spatially dependent functional data collected under a geostatistics setting, where locations are sampled from a spatial point process and a random function is observed at each location. Observations on each function are made on discrete time points and contaminated with measurement errors. The error process at each location is modeled as a non-stationary temporal process rather than white noise. Under the assumption of spatial isotropy, we propose a tensor product spline estimator for the spatio-temporal covariance function. If a coregionalization covariance structure is further assumed, we propose a new functional principal component analysis method that borrow information from neighboring functions. Under a unified framework for both sparse and dense functional data,
where the number of observations per curve is allowed to be of any rate relative to the number of functions, we develop the asymptotic convergence rates for the proposed estimators. The proposed methods are illustrated by simulation studies and a motivating example of the home price-rent ratio data in the New York metropolitan area.

Bayesian Spline Smoothing with Ambiguous Penalties
Xinlian Zhang, Gauri Datta, Ping Ma and Wenxuan Zhong
University of Georgia

A popular approach for flexible function estimation in nonparametric models is through spline smoothing using the general penalized likelihood method. In applying this method, one needs to specify a penalty functional which puts a soft constraint on the function to be estimated. A good choice of penalty functional is of key importance. In practice, specifying the penalty functional is mostly based on expert knowledge of the system. However, for many dynamic systems there naturally exists more than one sets of well-studied theories that explain the dynamics systems, i.e., there exist more than one sensible choices of penalties, rendering ambiguity in choosing a penalty form. To tackle this problem, we propose an approach that takes into consideration of all candidate penalties as well as the ambiguity in choosing among them. We take a fully Bayesian perspective, make use of the connection between penalized least squares and Bayesian estimation, and model the uncertainty of choosing penalty through introducing a mixture distribution as prior for parameters to be estimated. We also propose efficient sampling algorithm for making statistical inference based on taking samples from the posterior distribution.

Partially functional linear regression in high dimensions.
Hao Zhang1, Dehan Kong2, Fang Yao and Kaijie Xue
1University of Arizona
2University of Toronto

In modern experiments, functional and nonfunctional data are often encountered simultaneously when observations are sampled from random processes and high-dimensional scalar covariates. It is difficult to apply existing methods for model selection and estimation. We propose a new class of partially functional linear models to characterize the regression between a scalar response and covariates of both functional and scalar types. The new approach provides a unified and flexible framework that simultaneously takes into account multiple functional and ultrahigh-dimensional scalar predictors, enables us to identify important features, and offers improved interpretability of the estimators. The underlying processes of the functional predictors are considered to be infinite-dimensional, and one of our contributions is to characterize the effects of regularization on the resulting estimators. We establish the consistency and oracle properties of the proposed method under mild conditions, demonstrate its performance with simulation studies, and illustrate its application using air pollution data.

Session 66: SII Sponsored Invited Session: Modern Statistical Interface with Applications

Smoothing spline ANOVA for super-large samples: Scalable computation via rounding parameters
Nathaniel Helwig1 and Ping Ma2
1University of Minnesota
2University of Georgia

In the current era of big data, researchers routinely collect and analyze data of super-large sample sizes. Data-oriented statistical methods have been developed to extract information from super-large data. Smoothing spline ANOVA (SSANOVA) is a promising approach for extracting information from noisy data; however, the heavy computational cost of SSANOVA hinders its wide application. In this paper, we propose a new algorithm for fitting SSANOVA models to super-large sample data. In this algorithm, we introduce rounding parameters to make the computation scalable. To demonstrate the benefits of the rounding parameters, we present a simulation study and a real data example using electroencephalography data. Our results reveal that (using the rounding parameters) a researcher can fit nonparametric regression models to very large samples within a few seconds using a standard laptop or tablet computer.

Discovering Stock Chart Patterns by Statistical Estimation and Inference
Hoang Tran1 and Yiyuan She2
1Rho, Inc.
2Florida State University

Statistical modeling of stock price data is challenging due to heteroskedasticity, heavy-tails and outliers. These issues can be particularly relevant to the technical analysis practitioner who extracts trading signals from geometric patterns in prices. In this work, we propose a new method called Non-Parametric Outlier Identification and Smoothing (NOIS), which robustly smooths stock prices, automatically detects outliers and constructs pointwise confidence bands around the resulting curves. In real-world examples of high-frequency data, NOIS successfully detects erroneous prices as outliers and uncovers borderline cases for further study. NOIS can also highlight notable features and reveal new insights in inter-day chart patterns.

MCMC algorithms for empirical Bayes analysis of rank data
Vivekananda Roy1, Arnab Lah and Somak Dutta
1Iowa State University

Rank data arises frequently in marketing, finance, organizational behavior, and psychology. Most analysis of rank data reported in the literature assumes the presence of one or more variables (sometimes latent) based on whose values the items are ranked. We analyze rank data using a purely probabilistic model where the observed ranks are assumed to be perturbed versions of the true rank and each perturbation has a specific probability of occurring. We consider the general case when covariate information is present and has an impact on the rankings. An empirical Bayes approach is taken for estimating the model parameters. The Gibbs sampler is shown to converge very slowly to the target posterior distribution and we show that some of the widely used empirical convergence diagnostic tools may fail to detect this lack of convergence. We develop fast mixing MCMC algorithms for exploring the posterior distribution. An EM algorithm based on MCMC sampling is developed for estimating prior hyperparameters. A real life rank data set is analyzed using the proposed methods.

Dimension Reduction for Big Data
Tonglin Zhang
Purdue University

Dimension reduction is aimed at reducing the dimension of a high dimensional vector-valued explanatory variables and simultane
Automated Machine Learning

Jin Cao
Bell Labs, Nokia
jin.cao@nokia-bell-labs.com

AutoML is an new emerging research area that aims to develop algorithms to help non-machine learning experts to perform critical machine learning tasks that has been traditionally done by experts, for example, data preprocessing, feature learning and selection, model selection and hyper-parameter tuning. In this talk, we will present work conducted in Bell Labs Nokia in this endeavor. We shall focuses on two main areas: data structure discovery and feature learning. For data structure discovery, we present our methods in Identity Field discovery, periodicity detection for numeric fields, string pattern discovery and context inference. We also present methods for automatic feature learning for time series data, using two approaches, one using intelligent workflows based on expert statistical knowledge, and the other relying on deep learning. We will demonstrate our work using experimental results with public machine learning data repositories.

Application of Causal Bayesian Networks to Environmental Data

Jie Hu, Carlos Carrion, Andrew Mcgowan and Megan Chen
The Climate Corporation
hujie0704@gmail.com

In agriculture, we are often interested in studying how end-of-season crop yield responds to changes in soil nutrients measured during a growing season. For this purpose, it is necessary to identify the confounders in the soil nutrients-yield relationship because crop yield is the result of combined effects of soil nutrients, environment, management and weather. In addition, there are other potential problems with the data such as measurement variability due to different sampling times or sampling resolution. In this paper, we applied multiple constraint-based Causal Bayesian Networks Learning Algorithms to Climate Research Farm data from five farms spanning the U.S. corn belt to infer the structure of the relationships between the variables of interest and identify potential confounders. Moreover, we compared the inferred networks with domain knowledge in order to interpret results. We discovered several scenarios for which data suggested that yield responded to soil nutrients, as well as several scenarios where it did not. We identified potential confounders, which could be useful for correctly modeling the soil nutrients-yield relationship. Lastly, we also identified several aspects of the current Causal Bayesian Networks Learning Algorithms that can be improved.

Hierarchical Modeling and Shrinkage for User Session Length

Bias Correction For Paid Search In Media Mix Modeling

Aiyou Chen1, David Chan, Michael Perry, Yuxue Jin, Yunting Sun1, Nancy Wang and Jim Koehler
Google, LLC
aiyouchen@google.com

Evaluating the return on ad spend (ROAS), the causal effect of advertising on sales, is critical to advertisers for understanding the performance of their existing marketing strategy as well as how to improve and optimize it. Media Mix Modeling (MMM) has been used as a convenient analytical tool to address the problem using observational data. However it is well recognized that MMM suffers from various fundamental challenges: data collection, model specification and selection bias due to ad targeting, among others (Chan & Perry 2017; Wolfe 2016).

In this paper, we study the challenge associated with measuring the impact of search ads in MMM, namely the selection bias due to ad targeting. Using causal diagrams of the search ad environment, we derive a statistically principled method for bias correction based on the back-door criterion (Pearl 2013). We use case studies to show that the method provides promising results by comparison with results from randomized experiments. We also report a more complex case study where the advertiser had spent on more than a dozen media channels but results from a randomized experiment are not available. Both our theory and empirical studies suggest that in some common, practical scenarios, one may be able to obtain an approximately unbiased estimate of search ad ROAS.

Hierarchical Modeling and Shrinkage for User Session Length

We present a novel framework inspired by hierarchical Bayesian modeling to predict, at the moment of login, the amount of time a user will spend in the streaming service. The time spent by a user on a platform depends upon user-specific latent variables which are learned via hierarchical shrinkage. Our framework enjoys theoretical guarantees, naturally incorporates flexible parametric/nonparametric models on the covariates and is found to outperform state-of-the-art estimators in terms of efficiency and predictive performance on real world datasets.

Big-Data Analysis Points Toward A New Cancer Drug Discovery Method

Bin Chen
Michigan State University
bin.chen@hc.msu.edu

Rapidly decreasing costs of molecular measurement technologies not only enable profiling of disease sample molecular features (e.g., transcriptome, proteome, metabolome) at different levels (e.g., tissues, single cells), but also enable measuring of molecular signatures of individual drugs in clinically relevant models. Exploring methods to relate diseases to potentially efficacious drugs through various molecular features is critical in the discovery of new therapeutics. The target-based drug discovery approach that focuses on interfering with individual targets is challenged by the lack of drug efficacy, drug resistance, and off-target effects. We propose to employ a systems-based approach that identifies drugs that reverse the
molecular state of a disease. Using this system-based approach, we have successfully identified drug candidates for three cancers: Ewing’s sarcoma, liver cancer and basal cell carcinoma. In the Ewing’s sarcoma work, this systems-approach achieved a hit rate of > 50% in predicting effective drugs. In our recent integrative analysis of 66,000 compound gene expression profiles, 12 million compound activity measurements, 1,000 cancer cell line molecular profiles and 7,500 cancer patient samples, we uncovered that the drug’s ability to reverse cancer gene expression correlates to its efficacy, suggesting the potential of applying this approach in cancer drug discovery.

Mobile Health Apps Transforming Medical Research and Clinical Care: Using Apple’s new Research Kit for Asthma Mobile Health Study

Pei Wang

Icahn School of Medicine at Mount Sinai
pei.wang@mssm.edu

Asthma is one of the most common and costly of the chronic diseases, impacting a broad range of the population including both children and adults. It is a variable disease necessitating regular medication use, monitoring of symptoms, and avoidance of specific triggers. These characteristics of asthma make it a chronic disease particularly amenable to having a Mobile Health Application (MHA) facilitate active monitoring outside of periodic traditional medical visits. The asthma MHA is designed to facilitate asthma patient education and self-monitoring, promote positive behavioral changes, and reinforce adherence to treatment plans. Our research study assessed the association between the usage of the app and asthma symptom control, quality of life, and health care utilization. To date, the app has been downloaded approximately 50K times and over 8,000 individuals have electronically consented and enrolled in our Asthma Health research study. In the talk, we will describe the Asthma Mobile Health App as well as study findings and lessons learned to date. We will discuss the infrastructure, data management and data analysis we used. We will also discuss the benefits and limitations in conducting medical research with mobile devices, using Apple’s Research Kit and similar platforms as well as the promise of m-Health applications in improving clinical care.

Classification Errors and the Neyman-Pearson Classification

Jingyi Jessica Li

University of California, Los Angeles
jli@stat.ucla.edu

In this talk, I will first introduce the theoretic framework of binary classification and the concepts of population and empirical versions of the risk, the type I error and the type II error of a binary classifier, with a focus on their different sources of randomness. Second, I will introduce the Neyman-Pearson (NP) classification paradigm, which targets on minimizing the population type II error while enforcing an upper bound on the population type I error. Under the NP paradigm, I will further introduce our recent work including (1) an umbrella algorithm for implementing the NP classification with popular binary classification methods, such as Logistic Regression, Support Vector Machines, and Random Forests, (2) a graphical tool NP-ROC bands for visualizing NP classification results, and (3) a model selection criterion under the NP paradigm.

Session 69: Copula-based and mixed models to analyze next-generation sequencing data

Copula-based models for testing genetic effects in family studies: efficient sampling schemes and analysis

Richard Cook1 and Yujie Zong2
1University of Waterloo
2University of Cambridge

Studies on the genetic basis for disease are routinely conducted by selecting affected individuals called probands, and then recruiting, examining, and carrying out genetic tests of family members. Valid methods for assessing the genetic effects on disease risk must address the biased sampling scheme and the within-family dependence of the disease process. Mixed effect models are widely used to accommodate within-family dependence, but copula models offer a valid and appealing alternative approach which is particularly well-suited for estimating marginal effects of genetic markers. Within this framework we develop selection models for the sampling of families from a phase I sample of probands. Theses models may incorporate detailed information on the probands and summary data on their family members which are available at phase I. By exploiting this information we show that families may be chosen to yield more efficient marginal estimates of genetic effects on correlated times of disease onset, compared to those obtained by simple random sampling. The asymptotic efficiency gains are realized in empirical studies. An application to a motivating arthritis family study is given for illustration.

Multivariate linear mixed models for detecting gene-environment interactions

Hyonju Kim and Saunak Sen
University of Tennessee Health Science Center

Hyeonju Kim and Saunak Sen

We develop a multivariate linear mixed model with gene by environment (GxE) interactions that captures the correlation between phenotypes across multiple environments. Phenotypes are assumed to be correlated due to genetic similarity and environmental similarity. The model is estimated computationally efficiently taking advantage of structure of the covariance matrix. We apply our approach to a study of switchgrass yield in 11 sites across a latitudinal gradient in the United States.

Testing the heritability and parent-of-origin hypotheses for ages at onset of Psoriatic arthritis under biased sampling

Lajmi Lakhat-Chaieb1, Richard Cook2 and Yujie Zong3
1Laval University
2University of Waterloo
3University of Cambridge

The heritability and parent-of-origin effect hypotheses for chronic diseases can be evaluated by assessing the significance and strength of the parameters that measure the within-family dependences in disease onset times. We model the within-family associations in these times using a Gaussian copula whose correlation matrix accommodates the different pairwise family relationships. We derive score-type statistics to test the heritability and parent-of-origin effect hypotheses when the families selection protocol induces a sampling bias. We illustrate the use of the developed methods through an application to a motivating family study in Psoriatic arthritis and provide strong evidence of excessive paternal transmission of risk.
Regression analysis of dependent current status data using copulas
Ling Ma
Clemson University
lingm@clemson.edu
Current status data occurs when each subject is observed only once and the failure time of interest is either left- or right-censored. Many methods have been developed for the analysis of such data, most of which assume that the failure time and the observation time are independent or conditionally independent given covariates. This assumption may not hold in some real cases. For example, in tumorigenicity experiments, the failure time of interest is usually the time to tumor onset and the observation time is the death time (either natural death or sacrifice). The fact that most tumors are between lethal and nonlethal implies that the tumor onset and death times are most likely related. In such cases, we propose a sieve maximum likelihood approach with use of a copula model and monotone I-splines. The asymptotic properties of the resulting estimators are established. In particular, the estimated regression parameters are semiparametrically efficient.

Session 70: Biomarker and patient subgroup identification for precision medicine

Exhaustive Search for Subgroup Identification with Adaptive Binning and Screening
H. Zhang1 and J. Cheng2
1Merck & Co., Inc
2Abbvie Inc
hzhang@wpi.edu
As a core aspect of precision medicine, subgroup identification has enormous applications in clinical studies. The relationship between clinical outcomes and the patients’ baseline biomarkers is essential to develop personalized treatment to obtain better efficacy and less adverse events. We propose a subgroup identification method based on exhaustive search of differential treatment effect for subgroups defined by categorical covariates. A data-adaptive algorithm for binning and screening is also developed to account for continuous covariates. The proposed method can determine multivariate biomarker subgroup in terms of thresholds on individual biomarkers with well controlled type-I error by repeated cross-validation. Extensive simulation studies show that the proposed method has higher signal detection rate and biomarker discovery rate compared to a variety of existing methods. Real data analysis is also conducted to demonstrate the utility of the proposed method.

An exhaustive search based algorithm for patient subgroup identification
J. Cheng1 and H. Zhang
1Abbvie, Inc.
jie.cheng@abbvie.com
Central to precision medicine is the ability to effectively identify patient subgroups that have differential treatment effect of a certain treatment compared to the rest of the population. To identify patient subgroups, exploratory analysis is performed by searching a large search space of combinatorial biomarkers and clinical covariates. Due to severe multiple-testing, subgroup analysis is well known for its tendency to produce spurious results if not done with extreme care. In this talk, we will present an exhaustive search based subgroup identification method that controls false positives via random permutation and cross-validation techniques. The effectiveness of this method will be demonstrated using public clinical datasets.

Subgroups for precision medicine: from identification to confirmation
Lei Shen
Eli Lilly and Company
shen.lei@lilly.com
Over the last ten years, many new statistical methods have been developed to better identify patient subgroups with differential treatment effects from clinical trials during drug development. While the initial focus was to overcome the challenge of multiplicity inherent in a search among large number of candidate subgroups, a number of other aspects also play important roles as developers of new treatments move from identifying patient subgroups in early phase development to incorporating these subgroups in confirmatory trials. Two of these are estimation bias and level of uncertainty. The first can be (partially) addressed using resampling techniques, whereas the latter may best be characterized in a Bayesian framework. In this talk, I will attempt to make the connection between identification and confirmation of important patient subgroups (that is, those associated with differential treatment effects). Specifically, I will review the key information that should be provided by a desirable subgroup identification method, describe how such information impacts confirmatory testing during the next phase of development, and optimize the testing procedure to assess efficacy in multiple populations. Using some real examples and a simulation study, we hope to gain insight on the effective transition between subgroup identification and confirmation, thereby improving the overall effectiveness of the process of establishing clinically relevant subgroups toward precision medicine.

Session 71: New Developments in High Dimensional Data Analysis

High-dimensional statistical inferences with over-identification: confidence set estimation and specification test
Cheng Yong Tang1, Jinyuan Chang and Tong Tong Wu
1Temple University
yongtang@temple.edu
Over-identification is a signature feature of the influential Generalized Likelihood of Moments (Hansen, 1982) that flexibly allows more moment conditions than the Model parameters. Investigating over-identification together with high-dimensional statistical problems is challenging and remains less explored. In this paper, we study two high-dimensional statistical problems with over-identification. The first one concerns statistical inferences associated with multiple components of the high-dimensional model parameters, and the second one is on developing a specification test for assessing the validity of the over-identified moment conditions. For the first problem, we propose to construct a new set of estimating functions such that the impact from estimating the nuisance parameters becomes asymptotically negligible. Based on the new construction, a confidence set is estimated using empirical likelihood (EL) for the specified components of the model parameters. For the second problem, we propose a test statistic as the maximum of the marginal EL ratios respectively calculated from individual components of the high-dimensional moment conditions. Our theoretical analysis establishes the validity of the proposed procedures, accommodating exponentially growing data dimensionality, and our numerical examples demonstrate good performance and potential practical benefits of our proposed methods with high-dimensional problems.
Motivated by the co-heritability estimation problem in genomics, we study the inner product, quadratic functional and ratio in high-dimensional linear regression.

We observe \((X_i, Y_i)\) for \(N_1\) samples, and \((Z_j, W_j)\) for the other \(N_2\) samples. The feature vectors \(X_i\) and \(Z_j\) are observed on the same \(p\) features, and the response scalars \(Y_i\) and \(W_j\) are observed on different responses. The quantity of interest is the inner product of parameters vectors \(\theta_1\) for response 1 and \(\theta_2\) for response 2. The regime of interest is that \(\theta_1\) and \(\theta_2\) are sparse.

We propose a functional de-biased estimator for the inner product, quadratic functional and ratio. Beginning with the lasso estimator, we correct it by the estimation of the inner product between \(\theta_1\) (respectively, \(\theta_2\)) and error of the estimation for \(\theta_2\) (respectively, \(\theta_1\)), and take the summation as the corrected estimator. We prove that this method yields the optimal convergence rate, and show its competitive performance in numerical analysis. We also have the ready-to-use matlab package and R package for our methods.

Collaborative Spectral Clustering in Attributed Networks

Pengsheng Ji
University of Georgia

We propose a novel spectral clustering algorithm for attributed networks, where each node has \(p\)-dimensional meta-covariates from various formats such as text, image, speech, etc. The connectivity matrix \(W_{n \times n}\) is constructed with the adjacency matrix \(A_{n \times n}\) and covariate matrix \(X_{n \times p}\), and \(W = (1 - \alpha)A + \alpha K(X, X')\), where \(\alpha \in [0, 1]\) 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 show the consistent performance under both Stochastic Block Model (SBM) and Degree-Corrected Block Model (DCBM), especially in unbalanced networks where most community detection algorithms fail.

Graph-Based Two-Sample Tests for Discrete Data

Jingru Zhang and Hao Chen

In the regime of two-sample comparison, tests based on a graph constructed on observations by utilizing similarity information among them is gaining attention due to their flexibility and good performances under various settings for high-dimensional data and non-Euclidean data. However, when there are repeated observations or ties in terms of the similarity graph, these graph-based tests could be problematic as they are versatile to the choice of the similarity graph. We study two ways to fix the “tie” problem for the existing graph-based test statistics and a new max-type statistic. Analytic p-value approximations for these extended graph-based tests are also derived and shown to work well for finite samples, allowing the tests to be fast applicable to large datasets. The new tests are illustrated in the analysis of a phone-call network dataset. All proposed tests are implemented in R package gTests.
Differences in the Application of Dental Pain Model across Sites and Their Impact

Rina Leyva and Chunming Li
Pfizer Inc.

Dental pain is a well-accepted pain model which has a long and proven track record of discriminating between different analgesic drugs, different formulations and different doses of the same drugs. The two stopwatch techniques are also well validated measures of onset of analgesia commonly used in dental pain studies. Pfizer Consumer Healthcare (PCH) recently conducted two dental pain studies as part of a program development of a novel fixed dose combination product (FDC) of Ibuprofen (IBU) and Acetaminophen (APAP). This product is intended to provide superior analgesia compared to the maximum over-the-counter (OTC) doses of either monocomponent alone. Both studies were similar in design: double-blind, randomized trial in patients experiencing either moderate or severe post-operative pain within 5 hours of the end of oral surgery. The surgical and anesthesia techniques were standardized as were the assessment instruments to measure pain intensity, pain relief, time to onset, time to rescue and global evaluation. The two studies differed in that one was a full factorial single dose study conducted at Site 1 and the other was a multiple-dose study (FDC vs Placebo) conducted at Site 2. To measure analgesia both studies used time-weighted sum of Pain Intensity Difference (PID) scores (SPID) as the primary endpoint. Although, the results from both studies were successful in demonstrating the FDC was superior to placebo in terms of analgesia, the treatment differences measured by SPID were much lower for each of the treatment groups at Site 2 compared to those at Site 1. This prompted the team to investigate the reasons for the differences from a statistical perspective. Results were empirically reviewed including demographic/background, statistical assumptions, data handling, derivation of endpoints, data imputation and modeling. This review showed the two study populations were very similar but subjects at Site 2 were rescued earlier and more frequently which resulted in more missing data being imputed (via extrapolation method). This impacted the algorithm and outcome of pain differences and pain reliefs by shifting the scores toward more pain and less relief. It was also observed there were no differences in the onset of analgesia indicating subjects followed the instructions in the double stopwatch technique further validating this measure. The results of this data review were positively received at all levels in PCH clinical development as it provides an extra level of information for decision making in future dental pain studies. For example clinical operations can use this information during site selection or in the development of guidance documents. Clinicians and data management can make the decision to monitor data on how often and when subjects are being rescued. This work can also potentially help statisticians and stakeholders to formulate a framework in choosing estimands in dental or other different pain models. Within the current program, knowing the source in the treatment differences in these two studies will help drive the submission strategy as well as address potential FDA queries.

Sample Size Evaluation in Bioequivalence Studies with a Replicated Crossover Design

Jing Li and Chunming Li
Pfizer Inc.

In the design of bioequivalence (BE) studies, for highly variable drug products defined as those with a within-subject variability (i.e. %CV) equal or greater than 50% of the maximum concentration (Cmax) and/or the area under the concentration versus time curve (AUC), a reference-scaled BE approach using a replicate crossover design may be considered. The traditional method with acceptance range of 0.8 to 1.25 may no longer be applicable if such a replicate design is applied, depending on the regulatory requirements in the country or region of interest being considered. Different regulatory agencies have set up different guidelines in order to establish BE, including expanding the acceptance interval criteria. This poster presentation describes the sample size evaluation strategies by comparing the criteria of a couple of regulatory agencies, and presents the power calculations based on their guidelines via simulations using SAS.
ple have been trying to replace it with in vitro or other alternative methods (such as in vitro only, or in vitro + in vivo skin reflectance, which is non erythema SPF). Different artificial substrates close to skin’s physical characteristics have been used in the experiments. Because there is no intention to change the traditional definition of in vivo SPF which has been broadly used on sunscreen product labels, the goal of in vitro measurement is to show its equivalence with in vivo measurement. Review of the published in vitro SPF research reports shows that some statistical issues exist in the analyses and study designs, such as, superiority test was frequently used in many publications when the objective of these studies was to show the equivalence of the two methods. Even for equivalence tests, due to the differences in SPF measure, estimate and labeling by in vivo and in vitro, it is problematic to just compare head-to-head the SPF scores by simply using traditional equivalence tests. In this presentation, we will review these statistical issues and also the current equivalence testing methods. After introducing and comparing the differences in SPF measure and estimate by in vivo and in vitro methods, we will propose our testing procedures to compare their equivalence.

Session 74: The Future of Big Data and Artificial Intelligence

The Future of Big Data and Artificial Intelligence

Kelly Zou1, William Marler2, James Rosenberger3, Cheng Zhang4 and Tian Zheng5

1Pfizer Inc.
2IBM Watson Health
3National Institute of Statistical Sciences
4Clover Health
5Columbia University

Big data, real world data and data science have become ubiquitous and essential. Researchers, statisticians, and data scientists are challenged and tasked with leveraging evidence and gaining insights, generated from massive data. Various types of real world data, such as claims, electronic health records, surveys, and digital data, provide great opportunities for successful partnerships between academia and business, industry, and government organizations. This panel discussion session will highlight the opportunities to embrace and meet the challenge arising from big data, real world data and their related disciplines for the real world applications to gain insights. Specifically, the future of big data and artificial intelligence (AI) is discussed. To develop useful algorithms for the purpose of AI, it is important to understand data generating mechanisms and systems of data generation. We expect this session to have wide appeal, particularly given the increasing focus on interdisciplinary research and the emergence of complex data. The future generation of quantitative researchers and experts in data science and AI may be fostered. (Key Words: Artificial Intelligence, Big Data, Real World Data, Data Science, Partnership, Collaboration.)

Session 75: Statistical Analysis of Textual Data

Which Encoding is the Best for Text Classification in Chinese, English, Japanese and Korean?

Xiang Zhang

New York University

We offer an empirical study on the different ways of encoding Chinese, Japanese, Korean (CJK) and English languages for text classification. Different encoding levels are studied, including UTF-8 bytes, characters, words, romanized characters and romanized words. For all encoding levels, whenever applicable, we provide comparisons with linear models, fastText and convolutional networks. For convolutional networks, we compare between encoding mechanisms using character glyph images, one-hot (or one-of-n) encoding, and embedding. In total there are 473 models, using 14 large-scale text classification datasets in 4 languages including Chinese, English, Japanese and Korean. Some conclusions from these results include that byte-level one-hot encoding based on UTF-8 consistently produces competitive results for convolutional networks, that word-level n-grams linear models are competitive even without perfect word segmentation, and that fastText provides the best result using character-level n-gram encoding but can overfit when the features are overly rich.

Rare Feature Selection in High Dimensions

Xiaohan Yan1 and Jacob Bien2

1Cornell University
2University of Southern California

It is common in modern prediction problems for many predictor variables to be counts of rarely occurring events. This leads to design matrices in which many columns are highly sparse. The challenge posed by such “rare features” has received little attention despite its prevalence in diverse areas, ranging from natural language processing (e.g., rare words) to biology (e.g., rare species). We show, both theoretically and empirically, that not explicitly accounting for the rareness of features can greatly reduce the effectiveness of an analysis. We next propose a framework for aggregating rare features into denser features in a flexible manner that creates better predictors of the response. Our strategy leverages side information in the form of a tree that encodes feature similarity. We apply our method to data from TripAdvisor, in which we predict the number of reviews, and study designs, such as, superiority test was frequently used in many publications when the objective of these studies was to show the equivalence of the two methods. Even for equivalence tests, due to the differences in SPF measure, estimate and labeling by in vivo and in vitro, it is problematic to just compare head-to-head the SPF scores by simply using traditional equivalence tests. In this presentation, we will review these statistical issues and also the current equivalence testing methods. After introducing and comparing the differences in SPF measure and estimate by in vivo and in vitro methods, we will propose our testing procedures to compare their equivalence.

Organic ranking of competing domains

Junting Sun and Aiyou Chen

Google, LLC

Identification of competing domains given a target domain is often required in advertising and marketing research. A domain is considered as a competitor if it is competing for search traffic with the target domain. We propose a scalable algorithm to rank competing domains based on aggregated organic search data. The organic search data is in the form of a weighted multi-categorical bipartite graph, where the two sides of the graph represent domains and search queries. The accuracy of the ranking algorithm is evaluated using a proprietary ground truth data set based on survey.
Session 76: Student Session 2

Optimal calibration for computer model prediction with finite samples
Xiaowu Dai
University of Wisconsin-Madison
xda126@wisc.edu

We consider a non-asymptotic frequentist framework for computer model prediction. This framework concerns two main issues: (1) many computer models are inadequate for physical systems and (2) only finite samples of physical observations are available for estimating model discrepancy and calibrating multivariate unknown parameters in computer models. We propose a method to achieve the optimal calibration and provide exact statistical guarantees in the sense that the predictive mean squared error is minimized with optimal calibration for any finite samples. We derive an equivalent formulation of optimal calibration which leads naturally to an iterative algorithm. The connection is built between the optimal calibration and the Bayesian calibration in Kennedy and O’Hagan [J. R. Stat. Soc. Ser. B. Stat. Methodol. 63 (2001) 425-464]. Numerical simulations and a real data example show that the proposed calibration outperforms the existing ones in terms of the prediction.

Empirical-likelihood-based criteria for model selection on marginal analysis of longitudinal data with dropout missingness
Chixiang Chen and Ming Wang
Pennsylvania State University College of Medicine
chenxxy@psu.edu

Longitudinal data are common in clinical trials and observational studies, where the outcomes with missing data due to dropouts are always encountered. Under such context with the assumption of missing at random, weighted generalized estimating equations (WGEE) proposed by Robins et al. (1995) is widely adopted for marginal analysis. Of note is that model selection on marginal mean regression is a crucial aspect of data analysis, and identifying an appropriate correlation structure for model fitting may also be of interest. However, the existing information criteria for WGEE model selection have limitations, such as separate criteria for the selection of marginal mean regression and correlation structures, unsatisfactory selection performance in small-sample set-ups and so on. In particular, there exist few works developing joint information criteria for selection of both marginal mean and correlation structures. In this work, by embedding empirical likelihood into WGEE framework, we propose two innovative information criteria named joint empirical Akaike information criterion (JEAIC) and joint empirical Bayesian information criterion (JEBIC), which can simultaneously select the variables for marginal mean and correlation structures. Through extensive simulation studies, these empirical-likelihood-based criteria exhibit robustness, flexibility, and out-performance compared to other criteria including weighted quasi-likelihood under the independence model criterion (QICW), missing longitudinal information criterion (MLIC) and joint longitudinal information criterion (JLIC). More importantly, we provide rigorous theoretical justification of plug-in estimators in our proposed criteria and assess their asymptotic relationship with empirical likelihood estimates local spectra within blocks through penalized splines. CABS estimates parameters in computer models. We propose a method to achieve the optimal calibration and provide exact statistical guarantees in the sense that the predictive mean squared error is minimized with optimal calibration for any finite samples. We derive an equivalent formulation of optimal calibration which leads naturally to an iterative algorithm. The connection is built between the optimal calibration and the Bayesian calibration in Kennedy and O’Hagan [J. R. Stat. Soc. Ser. B. Stat. Methodol. 63 (2001) 425-464]. Numerical simulations and a real data example show that the proposed calibration outperforms the existing ones in terms of the prediction.

Online Sequential Leveraging Sampling Method for Streaming Time Series Data
Rui Xie, T. N. Sriram1, Wei Biao Wu and Ping Ma1
1University of Georgia
ruixie@uga.edu

Advances in data acquisition technology pose challenges in analyzing large volumes of streaming data. Sampling is a natural yet powerful tool for analyzing such data sets due to their competent estimation accuracy and low computational cost. Unfortunately, sampling methods and their statistical properties for streaming data, especially streaming time series data, are not well studied in the literature. In this article, we propose an online leverage-based sequential sampling algorithm for streaming time series data, which is assumed to come from an autoregressive model of order $p \geq 1$ (AR(p)). The proposed sequential leveraging sampling method samples only one consecutively recorded block from the data stream for inference. While the starting point of the sequential sampling scheme is chosen using a random mechanism based on leverage scores of the data, the subsample size is decided by a sequential sampling threshold. We show that an appropriately normalized sequential least squares estimator of the AR parameter vector is uniformly asymptotically normally distributed for non-explosive AR(p) model. Simulation studies and real data examples are presented to evaluate the empirical performance of the proposed sequential leveraging sampling method.

Session 77: Recent Advances of Statistical Modeling in Biomedical Research

Conditional adaptive Bayesian spectral analysis of nonstationary biomedical time series
Robert Krafty1, Scott Bruce2, Matricula Hall1 and Daniel Buysse1
1University of Pittsburgh
2George Mason University
rkrafty@pitt.edu

Many studies of biomedical time series signals aim to measure the association between frequency-domain properties of time series and clinical and behavioral covariates. However, the time-varying dynamics of these associations are largely ignored due to a lack of methods that can assess the changing nature of the relationship through time. We discuss a method for the simultaneous and automatic association of the time-varying power spectrum and covariates, which we refer to as conditional adaptive Bayesian spectral analysis (CABS). The procedure adaptively partitions the grid of time and covariate values into an unknown number of approximately stationary blocks and nonparametrically estimates local spectra within blocks through penalized splines. CABS is formulated in a fully Bayesian framework, in which the number and locations of partition points are random, and fit using reversible jump Markov chain Monte Carlo techniques. Estimation and inference averaged over the distribution of partitions allows for the accurate rate of convergence, higher smooth and abrupt changes. The proposed methodology is used to analyze the association between the time-varying spectrum of heart rate variability and self-reported sleep quality in a study of older adults serving as the primary caregiver for their ill spouse.

BAMM-SC: A Bayesian Mixture Model for Clustering Droplet-based Single Cell Transcriptomic Data from multiple individuals
Ying Ding1, Zhe Sun2, Ming Hu2 and Wei Chen1
1University of Pittsburgh
2Cleveland Clinic Lerner Research Institute
yingding@pitt.edu

Single cell transcriptome sequencing (scRNA-Seq) has become a revolutionary tool to study cellular and molecular processes at single cell resolution. Among existing technologies, the recently de-
Developed droplet-based platform enables efficient parallel processing of thousands of single cells with direct counting of transcript copies using Unique Molecular Identifier. We developed a novel Bayesian hierarchical Dirichlet multinomial mixture model, BAMM-SC, to cluster droplet-based scRNA-Seq data from multiple individuals simultaneously. To be noted, BAMM-SC is able to account for unbalanced sequencing depths among multiple individuals to reduce potential batch effect. We performed comprehensive simulations to evaluate BAMM-SC and compared it with existing clustering methods. In addition, we analyzed in-house scRNA-Seq datasets of peripheral blood mononuclear cells (PBMC) from healthy donors with prior biological knowledge to benchmark and validate BAMM-SC. Both simulation studies and real data applications demonstrated that overall, BAMM-SC achieves substantially improved clustering accuracy compared to other existing clustering methods.

Understanding Cellular Heterogeneity Using Single-cell Data
Lynn Lin
Pennsylvania State University
llin@psu.edu

Advances in single-cell technologies have enabled high-dimensional measurement of individual cells in a high-throughput manner. A key first step to analyze this wealth of data is to identify distinct cell subsets from a mixed-population sample. In many clinical applications, cell subsets of interest are often found in very low frequencies which pose challenges for existing clustering methods. To address this issue, we propose a new mixture model called Hidden Markov Model on Variable Blocks (HMM-VB) and a new mode search algorithm called Modal Baum-Welch (MBW) for mode-association clustering. HMM-VB leverages prior information about chain-like dependence among groups of variables to achieve the effect of dimension reduction as well as incisive modeling of the rare clusters. In case such a dependence structure is unknown or assumed merely for the sake of parsimonious modeling, we have developed a recursive search algorithm based on BIC to optimize the formation of ordered variable blocks. In addition, we provide theoretical investigations about the identifiability of HMM-VB as well as the consistency of our approach to search for the block partition of variables in a special case. In a series of experiments on simulated and real data, HMM-VB outperforms other widely used methods.

Fast and Robust Deconvolution of Tumor Infiltrating Lymphocyte from Expression Profiles using Least Trimmed Squares
Yuying Xie*, Ming Yan, Yu Lei and Yuning Hao
Michigan State University
xyy@msu.edu

Gene-expression deconvolution is a technique to quantify different types of cells in a mixed population. It holds enormous potential to characterize the immune landscape of solid cancers using whole tumor RNA-Seq datasets. However, a major challenge is that gene-expression data are usually contaminated with many outliers which decrease the estimation accuracy. It is imperative to develop a robust deconvolution method that automatically decontaminates data by reliably detecting and removing outliers. We developed a new machine learning tool, Fast And Robust DEconvolution of Expression Profiles (FARDEEP), to enumerate immune cell subsets from whole tumor tissue sequencing samples. To reduce noise in the tumor deep sequencing datasets, FARDEEP automatically detects and removes outliers from data before estimating the cell compositions. We show that FARDEEP is less sensitive to outliers and returns a better estimation of coefficients than the existing methods for both real data application and numerical simulations. In addition, FARDEEP provides the absolute quantitation of each immune cell subset in a mixed population.

Session 78: Statistical Inference in Air Pollution and Health Epidemiology

A Causal Inference Analysis of the Effect of Wildland Fire Smoke on Ambient Air Pollution Levels

Alexandra Larsen1, Shu Yang1, Ana Rappold2 and Brian Reich1
1North Carolina State University
2U.S. Environmental Protection Agency
aelarsen@ncsu.edu

The detrimental health outcomes associated with exposure to fine particulate matter, or PM2.5, are well-documented, as is the relationship between wildfire smoke and increased pollution levels. Our method separates wildfire-contributed PM2.5 from background concentrations in the contiguous U.S. using a causal inference framework and bias-adjusted computer simulations of PM2.5 under counterfactual scenarios. The numerical PM2.5 data for this analysis is from the Community Multi-Scale Air Quality (CMAQ) Modeling System, run with and without fire emissions across the contiguous U.S. for the 2008-2012 fire seasons. To account for biases, the CMAQ output is calibrated with observed data from the U.S. Environmental Protection Agency’s Federal Reference Method (FRM) PM2.5 monitoring sites for the same spatial domain and time period. We use a spatial Bayesian model to estimate the effect of wildfires on PM2.5 and state assumptions under which the estimate has a valid causal interpretation. The analysis is run over the U.S. partitioned into nine regions with similar climates. We find that the fire effect is highest in the areas of the U.S. where wildfires are most prevalent, including the West, Northwest and parts of the Southeast. Our results provide insight into using causal inference methods with numerical and spatial data as well as a method that can be extended to investigate the causal effects of wildfire smoke on mortality.

Efforts to Quantify the Causal Effect of Fine Particulate Matter on Mortality
Zhi Lin He, Zhengyuan Zhu and Hengfang Wang
Iowa State University
hezhulin@iastate.edu

Outdoor air pollution is a major environmental health problem affecting many countries. Sequential studies have indicated that higher PM2.5 exposure levels are associated with increases in mortality. However, it remains a challenging task to quantify the causal effect of PM2.5 on mortality. One difficulty is that PM2.5 exposure levels change over time and it is confounded by meteorological variables. Moreover, due to the existence of unmeasured confounder, the estimation of the effect of PM2.5 on mortality varies substantially from one study to another. In this study, we present a counterfactual approach to estimate the causal effect of time-varying PM2.5 exposure on mortality. In particular, by utilizing a directed acyclic graph description of the causal relationship, we propose a structural nested mean model under a spatial-temporal setting to estimate the causal effect of PM2.5.

Remove Confounding in Air Pollution Studies With Negative Controls
Wang Miao and Eric Tchetgen Tchetgen
Harvard University
mwfywxy@gmail.com

...
The observed associations between air pollution and public health outcomes are often confounded by unobserved variables, which lead to biased estimates of air pollution effects. However, in a timeseries study about the air pollution effects on health outcomes, we use future exposures and past outcomes as negative control variables to remove unmeasured confounding and to identify the causal effects. This is achieved under mild rank conditions, even if the confounder distribution may not be identified. We will also extend the negative control analysis into general settings.

A Bayesian critical window variable selection method for estimating the impact of air pollution exposure during pregnancy

Joshua Warren
Department of Biostatistics, Yale University
joshua.warren@yale.edu

Identifying periods of increased vulnerability during pregnancy to environmental exposures can improve our understanding of possible mechanisms of disease development and provide guidelines for protection of the fetus. Recently, advanced statistical methods have uncovered associations for a number of ambient air pollutants and adverse birth outcomes that traditional epidemiologic models have lacked the flexibility to detect. While these methods have improved our understanding of the importance of exposure timing during pregnancy, additional methods development is needed to overcome a number of analytic challenges. In this work, we develop a hierarchical Bayesian critical window variable selection model (CWVS) for estimating susceptible periods during pregnancy with respect to ambient air pollution exposure. CWVS improves upon existing methods by (i) offering a more accurate definition of a critical window of exposure, (ii) avoiding over-smoothing when estimating the risk magnitude, and (iii) extending to a logistic regression framework for more accessible odds ratio interpretations. We explore properties of the newly developed model through simulation and comparison with existing methods. We apply CWVS to birth records in North Carolina from 2005-2008 using ambient ozone and PM2.5 estimates from the EPA fused downscaler model with a focus on very preterm birth risk (i.e., 32 weeks of completed gestation). Simulation study results suggest that CWVS more often correctly identifies the true critical windows and improves estimation of the magnitude of risk compared to existing methods. Data application results suggest an association between increased ozone exposure during the first trimester and an increased risk of very preterm birth.

Causal inference in the context of an error prone exposure: air pollution and mortality

Xiao Wu¹, Danielle Braun², Marianthi-Anna Kiourmourtzoglou³, Christine Choirat⁴, Qian Di⁵ and Francesca Dominici²
¹Harvard University
²Harvard T.H. Chan School of Public Health
³Mailman School of Public Health, Columbia University
wuaxiao@g.harvard.edu

We propose a new approach for estimating causal effects when the exposure is measured with error and confounding adjustment is performed via a generalized propensity score (GPS). Using validation data, we propose a regression calibration (RC)-based adjustment for a continuous error-prone exposure combined with GPS to adjust for confounding (RC-GPS). The outcome analysis is conducted after transforming the corrected continuous exposure into a categorical exposure. We consider confounding adjustment in the context of GPS sub-classification, inverse probability treatment weighting (IPTW) and matching. In simulations with varying degrees of exposure error and confounding bias, RC-GPS eliminates bias from exposure error and confounding compared to standard approaches that rely on the error-prone exposure. We applied RC-GPS to a rich data platform for estimating the causal effect of long-term exposure to fine particles (PM2.5) on mortality in New England for the period from 2000 to 2012. The main study consists of 2,202 zip codes covered by 217,660 1km² grid cells with yearly mortality rates, yearly PM2.5 averages estimated from a spatio-temporal model (error prone exposure) and several potential confounders. The internal validation study includes a subset of 83 1km² grid cells within 75 zip codes from the main study with error-free yearly PM2.5 exposures obtained from monitor stations. Under assumptions of non-interference and weak unconfoundedness, using matching we found that exposure to moderate levels of PM2.5 (8 < PM2.5 ≤ 10 µg/m³) causes a 2.8% (95% CI: 0.6%, 3.6%) increase in all-cause mortality compared to low exposure (PM2.5 < 8 µg/m³).

Session 79: New clinical research strategies in oncology

Basket Trial Design in Oncology Studies
Kun Nie
Clindata Insight, Inc.
knie@clindatainsight.com

In the evaluation of targeted therapies, basket trials have emerged as an approach to test the hypothesis that targeted therapies may be effective independent of tumor histology, as long as the molecular target is present. A big advantage of the basket design is that the question of efficacy of a targeted agent can be assessed with fewer patients and in a shorter amount of time when compared with the traditional trial design because the data from different arms that show similar efficacy at the interim analysis can be aggregated. This enables early termination of arms not likely to show efficacy. In this presentation, we will propose a basket design that requires smallest number of patients in order to observe either pre-specified number of responders or non-responders, and each baskets can be closed whenever efficacy or futility can be claimed. Bayesian approach that pool data from different baskets will also be discussed.

Statistical learning for biomarker discovery and development in cancer prevention studies
Xi Zhou
Weill Cornell Medical College
kaz2004@med.cornell.edu

In recent years, biomarker discovery and development have become an important component of cancer prevention research with the goal of improved understanding of premalignant biology for more precise risk assessment and targeted early intervention. Despite the plethora of statistical learning methods and tools, challenges often arise in practical applications. In this talk, I will discuss our work related to biomarker discovery and development in the cancer prevention setting. In the first part of the talk, I will discuss the development of a Bayesian model averaging approach for improved identification of differentially expressed genes from heterogeneous observational samples. In the second part of the talk, I will discuss our work on developing biomarkers predictive of the presence of crown like structure, a lesion of subclinical inflammation in adipose tissues, in two fat depots using blood biomarkers. The predictive biomarkers may allow improved patient selection for cancer prevention trials.

Penalized multiple inflated values selection method with application to SAFER data
Yang Li
Multiple Testing for temporal pattern change detection and identification
Zhi Wei
New Jersey Institute of Technology
zhiwei@njit.edu

Time course experiments are generally conducted to capture temporal changes. It is generally interested to detect if any changes happen over time, which we define as a detection problem. If there is a change, it is informative to know where/when the change happens, which we define as an identification problem. Power for the detection problem is generally higher than the identification problem, while knowing the change-point location is more desirable than just knowing there is a change. Quite a few methods have been developed for analysis of time course data. However, most of existing methods aim to solve either the detection problem or, more recently, the identification problem. Here, we propose to solve these two problems in a unified multiple-testing framework built upon change-point models. Users are given the flexibility to solve the detection or identification problems. We show that the proposed approach is valid in terms of controlling type I error and is optimal in terms of power. Moreover, a specially designed simulation, based on the structure of data from a randomized clinical trial of an HIV sexual risk education intervention, performs well and ensures our method could be generalized to the real situation. The empirical analysis of a clinical trial dataset is used to elucidate the MIP model.

Session 80: Addressing the issue of multiplicity and selection in modern data analysis

Multiple Testing for temporal pattern change detection and identification
Zhi Wei
New Jersey Institute of Technology
zhiwei@njit.edu

Time course experiments are generally conducted to capture temporal changes. It is generally interested to detect if any changes happen over time, which we define as a detection problem. If there is a change, it is informative to know where/when the change happens, which we define as an identification problem. Power for the detection problem is generally higher than the identification problem, while knowing the change-point location is more desirable than just knowing there is a change. Quite a few methods have been developed for analysis of time course data. However, most of existing methods aim to solve either the detection problem or, more recently, the identification problem. Here, we propose to solve these two problems in a unified multiple-testing framework built upon change-point models. Users are given the flexibility to solve the detection or identification problems. We show that the proposed approach is valid in terms of controlling type I error and is optimal in terms of power. Moreover, a specially designed simulation, based on the structure of data from a randomized clinical trial of an HIV sexual risk education intervention, performs well and ensures our method could be generalized to the real situation. The empirical analysis of a clinical trial dataset is used to elucidate the MIP model.

Estimation with truncated data
Asaf Weinstein and Amit Meir
Stanford University
asafw@stanford.edu

We consider a “selective” version of the classical normal means problem, where one observes only measurements that exceed a threshold, and needs to estimate the underlying means under sum-of-squares loss. We propose an empirical Bayes method, through which we demonstrate that, as in the un-truncated problem, the gain from pooling can be substantial.

Local False Discovery Rate Based Methods for Multiple Testing of One-Way Classified Hypotheses
Zhigen Zhao and Sanat Sarkar
Temple University
Zhigen.zhao@gmail.com

This paper continues the line of research initiated in Liu, Sarkar and Zhao (2016) on developing a novel framework for multiple testing of hypotheses grouped in a one-way classified form using hypothesis-specific local false discovery rates (Lfdr’s). It is built on an extension of the standard two-class mixture model from single to multiple groups, defining hypothesis-specific Lfdr as a function of the conditional Lfdr for the hypothesis given that it is within a significant group and the Lfdr for the group itself and involving a new parameter that measures grouping effect. This definition captures the underlying group structure for the hypotheses belonging to a group more effectively than the standard two-class mixture model. Two new Lfdr based methods, possessing meaningful optimalities, are produced in their oracle forms. One, designed to control false discoveries across the entire collection of hypotheses, is proposed as a powerful alternative to simply pooling all the hypotheses into a single group and using commonly used Lfdr based method under the standard single-group two-class mixture model. The other is proposed as an Lfdr analog of the method of Benjamini and Bogomolov (2014) for selective inference. It controls Lfdr based measure of false discoveries associated with selecting groups concurrently with controlling the average of within-group false discovery proportions across the selected groups. Numerical studies show that our proposed methods are indeed more powerful than their relevant competitors, at least in their oracle forms, in commonly occurring practical scenarios.

Session 81: New methods with modern large and complex data sets

Pre-processing with Orthogonal Decompositions for High-Dimensional Explanatory Variables
Xu Han
Temple University
hanxu3@temple.edu

It is well known that high level of correlations between explanatory variables is problematic for high-dimensional regularized regression methods targeting at estimating sparse linear models. Due to the violation of the irrepresentable condition, the popular LASSO method may suffer from false inclusions of non-contributing variables. The penalized estimation approaches may suffer from false inclusions of non-contributing variables when encountering highly correlated explanatory variables. In this paper, we propose pre-processing with orthogonal decompositions (PROD) for the explanatory variables in high-dimensional regressions. The PROD procedure is constructed based upon a generic orthogonal decomposition of the design matrix. We investigate in detail three specific cases of the PROD: one by the conventional principal component analysis, one by a novel optimization incorporating the impact from the response variable, and one by random projections. We show that the level of correlations can be effectively reduced with PROD, making it more realistic for the irrepresentable condition to be valid. We also recognize that the PROD is flexible and can be adapted taking multiple objectives into consideration such as reducing the level of correlations between the explanatory variables yet without compromising the level of variations associated with the resulting estimator. Extensive numerical studies with simulations and data analysis show the promising performance of the PROD. Our theoretical analysis also confirms its effect and benefit for high-dimensional regularized regression methods.

Multiple change point detection for manifold-valued data with...
applications to dynamic functional connectivity

Qiang Sun
University of Toronto
qsun.ustc@gmail.com

In neuroscience, functional connectivity describes the connectivity between brain regions that share functional properties. It is often characterized by a time series of covariance matrices between functional measurements of distributed neuron areas. An effective statistical model for functional connectivity and its changes over time is critical for better understanding neurological diseases. To this end, we propose a log-mean model with an additive heterogeneous noise for modeling random symmetric positive definite matrices that lie in a Riemannian manifold. The heterogeneity of error terms is introduced specifically to capture the curved nature of the manifold. A scan statistic is then developed for the purpose of multiple change point detection. Despite that the proposed model is linear and additive, it is able to account for the curved nature of the Riemannian manifold. Theoretically, we establish the sure coverage property. Simulation studies and an application to the Human Connectome Project lend further support to the proposed methodology.

Empirical Band Analysis of Nonstationary Time Series

Scott Bruce
George Mason University
sbruce@gmu.edu

Power spectra of time series processes are defined over a continuous range of frequencies. However, time series data contain only a finite number of observations, so we must consider collapsed measures of power within local frequency bands that partition the frequency space. Frequency bands are used widely in the scientific literature and are often selected by manual observation of waveforms generated from a specific type of signal under particular settings. This article provides a standardized, unifying approach to constructing customized frequency bands for different signals under study across different settings. A frequency-domain, iterative cumulative sum procedure is formulated to identify optimal frequency bands that best preserve nonstationary information. A formal hypothesis testing procedure is also developed to test which, if any, frequency bands remain stationary. This method is shown to consistently estimate the number of frequency bands and the location of the upper and lower bounds defining each frequency band.

A Bayesian General Linear Modeling Approach to Cortical Surface fMRI Data Analysis

Yu Yue
Baruch College, The City University of New York
yu.yue@baruch.cuny.edu

Cortical surface fMRI (cs-fMRI) has recently gained in popularity versus traditional volumetric fMRI, as it allows for more meaningful spatial smoothing and is more compatible with the common assumptions of isotropy and stationarity in Bayesian spatial models. However, as no Bayesian spatial model has been proposed for cs-fMRI data, most analyses continue to employ the classical, voxel-wise general linear model (GLM) (Worsley and Friston 1995). Here, we propose a Bayesian GLM for cs-fMRI, which employs a class of sophisticated spatial processes to flexibly model latent activation fields. We use integrated nested Laplacian approximation (INLA), a highly accurate and efficient Bayesian computation technique (Rue et al. 2009). To identify regions of activation, we propose an excursions set method based on the joint posterior distribution of the latent fields, which eliminates the need for multiple comparisons correction. Finally, we address a gap in the existing literature by proposing a novel Bayesian approach for multi-subject analysis. The methods are validated and compared to the classical GLM through simulation studies and a motor task fMRI study from the Human Connectome Project. The proposed Bayesian approach results in smoother activation estimates, more accurate false positive control, and increased power to detect truly active regions.

Session 82: Recent Advances in Genomic Data Analysis

Pedigree-based Functional Linear Mixed Models for Association Analysis of Quantitative Traits with Next-Generation Sequencing Data

*Rucong Fan*, Bingsong Zhang, Ao Yuan, Chi-Yang Chiu, Fang Yuan, Xin Li, Hong-Bin Fang, Kenneth Lange, Alexander Wilson, Joan Bailey-Wilson, M. hamed Lajmi, Lahal-Chaiteb, Richard Cook, Francis Mcmahon, Christopher Amos and Momiao Xiong

1. Georgetown University Medical Center
2. University of Tennessee Health Science Center
3. Kunning Medical University
4. University of California, Los Angeles
5. University of Pittsburgh
6. National Institutes of Health
7. Université Laval
8. University of Waterloo
9. Dartmouth Medical School
10. University of Texas at Houston

In this project, we develop functional linear mixed models (FLMM) to analyze quantitative traits of pedigree data. In the models, the effect of a major gene modeled as a fixed mean function, the contributions of polygenes and major gene are modeled as two separate random variations, and the correlation of pedigree members is modeled by inbreeding/kinship coefficients and major gene proportions of alleles shared identical by state (IBS). In addition, additive linear mixed models (LMM) are presented for a full investigation and comparison. Likelihood ratio test (LRT) statistics based on the additive LMM and FLMM are built to test for association between a quantitative trait and the genetic variants. By simulation, show that the LRT statistics of FLMM control type I errors correctly. The proposed models are useful in whole genome and whole exome association studies of complex traits.

Gene-based Association Testing of Dichotomous Traits with Generalized Linear Mixed Models Using Extended Pedigrees

*Chi-Yang Chiu*, Yingda Jiang, Daniel Weeks and Rucong Fan

1. The University of Tennessee Health Science Center
2. University of Pittsburgh
3. Georgetown University Medical Center

We propose two approaches to test association between a dichotomous trait and genetic variants in a chromosome region for family-based data. The two approaches are based on additive generalized linear mixed models (GLMM) and generalized functional linear mixed models (GFLMM). The GLMM and GFLMM model the effect of a major gene as a fixed mean, the polygenic contributions as a random variation, and the correlation of pedigree members by kinship coefficients. The association between the dichotomous trait and the genetic variants is tested by testing the fixed mean to be 0 using likelihood ratio test (LRT) statistics. Simulation results indicate that the proposed LRT statistics control the type I error rates accurately and have higher power than the retrospective kernel and bur-
den statistics developed by Schaid and his colleagues in most scenarios. The applications of our proposed methods to an age-related macular degeneration (AMD) family data set detect strong association between AMD and two known AMD susceptibility genes: CFH and ARMS2.

Meta-analytic and integrative framework for sparse K-means to identify disease subtypes (Zhi guang Hao and George Tseng)
1Department of Biostatistics, University of Florida
2University of Pittsburgh
zhao@ufl.edu

Disease phenotyping by omics data has become a popular approach that potentially can lead to better personalized treatment. Identifying disease subtypes via unsupervised machine learning is the first step towards this goal. With the accumulation of massive high-throughput omics data sets, omics data integration is essential to improve statistical power and reproducibility. In this talk, two extensions of sparse K-means method will be introduced. The first extension is towards a meta-analytic framework to identify novel disease subtypes when expression profiles of multiple cohorts are available. The lasso regularization and meta-analysis identify a unique set of gene features for subtype characterization. An additional pattern matching reward function guarantees consistent subtype signatures across studies. The second extension is towards integrating multi-level omics datasets with the guidance of prior biological knowledge via sparse overlapping group lasso. An algorithm using alternating direction method of multiplier (ADMM) will be applied for fast optimization. Simulation and real applications in breast cancer and leukemia will show the superior clustering accuracy, feature selection, functional annotation and computing efficiency.

RNASeqDesign: A framework for RNA-Seq genome-wide power calculation and study design issues (Chien-Wei Lin, George Tseng, Yongseok Park, Ge Xiao and Mei-Ling Ting Lee)
1Medical College of Wisconsin
2University of Pittsburgh
3University of Maryland, College Park
chlin@mcw.edu

Massively parallel sequencing (a.k.a. next-generation sequencing, NGS) technology has emerged as a powerful tool in characterizing genomic profiles. Among many NGS applications, RNA sequencing (RNA-Seq) has gradually become a standard tool for global transcriptomic monitoring. Although the cost of NGS experiments has dropped constantly, the high sequencing cost and bioinformatic complexity are still obstacles for many biomedical projects. Unlike earlier fluorescence-based technologies such as microarray, modeling of NGS data should consider discrete count data. In addition to sample size, sequencing depth also directly relates to the experimental cost. Consequently, given total budget and pre-specified unit experimental cost, the study design issue in RNA-Seq is conceptually a more complex multi-dimensional constrained optimization problem rather than one-dimensional sample size calculation in traditional hypothesis setting. In this paper, we propose a statistical framework, namely “RNASeqDesign”, to utilize pilot data for power calculation and study design of RNA-Seq experiments. The approach is based on mixture model fitting of p-value distribution from pilot data and a parametric bootstrap procedure based on approximated Wald test statistics to infer genome-wide power for optimal sample size and sequencing depth. We further illustrate five practical study design tasks for practitioners. We perform simulations and three real applications to evaluate the performance and compare to existing methods.

Session 83: Statistical Learning for Complex and High-Dimensional Data with Modern Applications

Blessing of Massive Scale: Total Cardinality Approach for Spatial Graphical Model Estimation (Ethan Fang)
Pennsylvania State University
xxfl@psu.edu

We consider the problem of estimating high dimensional spatial graphical models with a total cardinality constraint (i.e., the L0-constraint). Though this problem is highly nonconvex, we show that its primal-dual gap diminishes linearly with the dimensionality and provide a convex geometry justification of this “blessing of massive scale” phenomenon. Motivated by this result, we propose an efficient algorithm to solve the dual problem (which is concave) and prove that the solution achieves optimal statistical properties. Extensive numerical results are also provided.

Semi-supervised Inference for Explained Variance in High-dimensional Linear Regression and Its Applications (Zijian Guo and Tony Cai)
1Rutgers University
2University of Pennsylvania
zijguo@stat.rutgers.edu

We consider statistical inference for the explained variance $\beta^T\Sigma\beta$ under the high-dimensional linear model $Y = X\beta + \epsilon$ in the semi-supervised setting, where $\beta$ is the regression vector and $\Sigma$ is the design covariance matrix. A calibrated estimator, which efficiently integrates both labelled and unlabelled data, is proposed. It is shown that the estimator achieves the minimax optimal rate of convergence in the general semi-supervised framework. The optimality result characterizes how the unlabelled data affects the minimax optimal rate. Moreover, the limiting distribution for the proposed estimator is established and data-driven confidence intervals for the explained variance are constructed. We further develop a randomized calibration technique for statistical inference in the presence of weak signals and apply the obtained inference results to a range of important statistical problems, including signal detection and global testing, prediction accuracy evaluation, and confidence band construction. The numerical performance of the proposed methodology is demonstrated in simulation studies and an analysis of estimating heritability for a yeast segregant data set with multiple traits.

Learning causal networks via additive faithfulness (Kuang-Yao Lee, Tianqi Liu, Bing Li and Hongyu Zhao)
1Temple University
2Kenso
3Pennsylvania State University
4Yale University
kuangyao.l@gmail.com

In this work we introduce a statistical model, called additively faithful directed acyclic graph (AFDAG), for causal learning from observational data. Our approach is based on additive conditional independence (ACI), a recently proposed three-way statistical relation that shares many similarities with conditional independence but without resorting to multivariate kernels. This distinct feature strikes a balance between a parametric model and a fully nonparametric model, which makes the proposed model attractive to large networks. For graph inference, we develop an estimator for AFDAG...
based on a linear operator that characterizes ACI, and establish the consistency and convergence rates of our estimator. Moreover, we prove the uniform consistency of the estimated DAG under a stronger additive faithfulness condition, which appears to be less restrictive than its linear counterpart. We introduce a modified PC-algorithm to implement the estimating procedures efficiently, so that their complexity is determined by the level of sparsity rather than the dimension of the network. Through simulation studies we show that our method outperforms existing methods when commonly assumed conditions such as Gaussian or Gaussian copula distributions do not hold. Finally, the usefulness of AFDAG formulation is demonstrated through an application to a proteomics data set.

Minimax Estimation of Large Precision Matrices with Bandable Cholesky Factor
Zhao Ren
University of Pittsburgh
zren@pitt.edu

Last decade witnesses significant methodological and theoretical advances in estimating large precision matrices. In particular, there are scientific applications such as longitudinal data, meteorology and spectroscopy in which the ordering of the variables can be interpreted through a bandable structure on the Cholesky factor of the precision matrix. However, the minimax theory has still been largely unknown, as opposed to the well established minimax results over the corresponding bandable covariance matrices. In this paper, we focus on two commonly used types of parameter spaces, and develop the optimal rates of convergence under both the operator norm and the Frobenius norm. A striking phenomenon is found: two types of parameter spaces are fundamentally different under the operator norm but enjoy the same rate optimality under the Frobenius norm, which is in sharp contrast to the equivalence of corresponding two types of bandable covariance matrices under both norms. This fundamental difference is established by carefully constructing the corresponding minimax lower bounds. Two new estimation procedures are developed: for the operator norm, our optimal procedure is based on a novel local cropping estimator targeting on all principle submatrices of the precision matrix while for the Frobenius norm, our optimal procedure relies on a delicate regression-based block-thresholding rule. Adaptive estimation is achieved using Lepski’s method. We further establish rate optimality in the nonparanormal model. Numerical studies are carried out to confirm our theoretical findings.

Session 84: Recent advances in high-dimensional statistics

Adaptive QDA to dimensionality
Qing Yang1 and Guang Cheng2
1Department of Statistics, Purdue University
2Purdue University
yang1512@purdue.edu

Quadratic discriminant analysis (QDA) is a well-known method to classify a subject into two populations with different covariance matrices. In the classical setting when the data dimension p is fixed, the QDA has the smallest asymptotic misclassification rate. However, when p is comparable with the sample sizes, this nice property disappears and it can perform as poorly as random guessing even when the two populations deviate enough. This motivates us to propose an adaptive QDA to dimensionality such that its asymptotic misclassification rate is close to the optimal one when the two populations deviate sufficiently. In the more difficult scenario of two close populations, we further apply a divide-and-conquer strategy to enforce the rate to be arbitrarily close to the optimal one.

Robust estimation, efficiency, and Lasso debiasing
Po-Ling Loh
University of Wisconsin-Madison
polingloh@gmail.com

We present results concerning high-dimensional robust estimation for linear regression with non-Gaussian errors. We provide error bounds for certain local/global optima of penalized M-estimators, valid even when the loss function employed is nonconvex – giving rise to more robust estimation procedures. We also present a new approach for robust location/scale estimation with rigorous theoretical guarantees. We conclude by discussing high-dimensional variants of one-step estimation procedures from classical robust statistics and connections to recent work on confidence intervals based on Lasso debiasing.

Learning Nonconvex Hierarchical Interactions
Xingzhong Xue1 and Hongbo Dong
1Pennsylvania State University
lxzxe@psu.edu

In this talk, we will focus on learning nonconvex hierarchical interactions in high-dimensional statistical models. We first use the affine sparsity constraints to provide a precise characterization of both strong and weak hierarchical interactions. However, these affine sparsity constraints do not lead to a closed feasible region. To address this issue, we derive the explicit closure of the affine sparsity constraint for learning nonconvex hierarchical interactions. We prove that the global solution can be found by solving a sequence of folded concave penalized estimation and the desired strong or weak hierarchy holds with probability one. Furthermore, we study the local convergence theory for learning hierarchical interactions using the folded concave penalized estimation. Numerical studies are used to demonstrate the power of our proposed methods.

Randomized incomplete U-statistics in high dimensions
Xiaohui Chen1 and Kengo Kato2
1University of Illinois at Urbana-Champaign
2University of Tokyo
xchen@illinois.edu

This paper studies inference for the mean vector of a high-dimensional U-statistic. In the era of Big Data, the dimension d of the U-statistic and the sample size n of the observations tend to be both large, and the computation of the U-statistic is prohibitively demanding. Data-dependent inferential procedures such as the empirical bootstrap for U-statistics is even more computationally expensive. To overcome such computational bottleneck, incomplete U-statistics obtained by sampling fewer terms of the U-statistic are attractive alternatives. In this paper, we introduce randomized incomplete U-statistics with sparse weights whose computational cost can be made independent of the order of the U-statistic. We derive non-asymptotic Gaussian approximation error bounds for the randomized incomplete U-statistics in high dimensions, namely in cases where the dimension d is possibly much larger than the sample size n, for both non-degenerate and degenerate kernels. In addition, we propose novel and generic bootstrap methods for the incomplete U-statistics that are computationally much less-demanding than existing bootstrap methods, and establish finite sample validity of the proposed bootstrap methods. The proposed bootstrap methods are illustrated on the application to nonparametric testing for the pairwise independence of a high-dimensional random vector under...
weaker assumptions than those appearing in the literature.

Session 85: Nonparametric and high dimensional statistics

Mean Field Variational Inference: Computational and Statistical Guarantees
Anderson Zhang
Yale University
ye.zhang@yale.edu
Mean field variational inference is widely used in statistics and machine learning to approximate posterior distributions. Despite its popularity, there exist remarkably little fundamental theoretical justifications. The success of variational inference mainly lies in its iterative algorithm, which, to the best of our knowledge, has never been investigated for any high-dimensional or complex model. In this talk, I will describe the statistics/computation interface of the iterative algorithm of mean field variational inference. I will study it from a frequentist perspective, quantifying it by posterior contraction. For community detection problem, I will show that the iterative algorithm has a linear convergence to the optimal statistical accuracy within log n iterations.

The technique can be extended to analyzing Expectation-maximization and Gibbs sampler with similar guarantees obtained, which I will briefly describe. The community detection problem used in my talk provides a test case and playground, and it is promising to understand mean field under a general class of latent variable models.

Trace regression with nonconvex regularization
Tingni Sun
University of Maryland, College Park
tingni@math.umd.edu
Statistical estimation of low-rank matrices has recently attracted increasing attention. We consider the generalized trace regression model and estimate the coefficient matrix via the empirical negative log-likelihood minimization with the nonconvex penalty on the matrix spectrum. The property of nonconvex penalty is studied in the matrix version and leads to non-asymptotic error bounds under the Frobenius norm for estimating the coefficient matrix. A version of proximal gradient method is applied to efficiently solve the nonconvex minimization problem. An analysis of personality assessment datasets illustrates the superior performance of using nonconvex regularization.

BET on Independence
Kai Zhang
University of North Carolina at Chapel Hill
zhangk@email.unc.edu
We study the problem of nonparametric dependence detection. Many existing methods suffer severe power loss due to non-uniform consistency, which we illustrate with a paradox. To avoid such power loss, we approach the nonparametric test of independence through the new framework of binary expansion statistics (BEStat) and binary expansion testing (BET), which examine dependence through a novel binary expansion filtration approximation of the copula. Through a Hadamard transform, we find that the cross interactions of binary variables in the filtration are complete sufficient statistics for dependence. These interactions are also uncorrelated under the null. By utilizing these interactions, the BET avoids the problem of non-uniform consistency and improves upon a wide class of commonly used methods (a) by achieving the minimax rate in sample size requirement for reliable power and (b) by providing clear interpretations of global relationships upon rejection of independence. The binary expansion approach also connects the test statistics with the current computing system to facilitate efficient bitwise implementation. We illustrate the BET with a study of the distribution of stars in the night sky and with an exploratory data analysis of the TCGA breast cancer data.

Maximum likelihood estimation of Sublinear Preferential Attachment Models and its connection to Janson’s urn models
• Fengnan Gao1 and Aad Van der Vaart
1Fudan University
fngao@fudan.edu.cn
The preferential attachment (PA) network is a popular way of modeling the social networks, the collaboration networks and etc. The PA network model is an evolving network model where new nodes keep coming in. When a new node comes in, it establishes only one connection with an existing node. The random choice on the existing node is via a multinomial distribution with probability weights based on a preferential function f on the degrees. f maps the natural numbers to the positive real line and is assumed apriori non-decreasing, which means the nodes with high degrees are more likely to get new connections, i.e., “the rich get richer”. Under sub-linear parametric assumptions on the PA function with parameter \( \theta \in \Theta \), we proposed the maximum likelihood estimator on \( \theta \). We show that the MLE yields optimal performance with the asymptotic normality results. The only drawback of the MLE is its dependence on the history of the network evolution, which is often difficult to obtain in practice. To avoid such shortcomings of the MLE, we propose the quasi maximum likelihood estimator (QMLE), a history-free remedy of the MLE. To prove the asymptotic normality of the QMLE, the connection between the PA model and Svante Janson’s urn models is exploited.

Session 86: Leveraging Evidence from Historical/Subsidiary Data in Randomized Controlled Trial

The use of alternative data in clinical trials in small population and pediatric patients
Shiling Ruan
Novartis
shilingr@hotmail.com
In some disease areas, such as rare disease or pediatric disease, it could be quite challenging to conduct a randomized clinical trial due to difficulty in patient enrollment or other practical concerns. On the other hand, in these areas, there are often valuable alternative data available from different sources, such as historical studies, published reports, natural history studies, or concurrent studies. If relevant and reliable alternative data are available, with appropriate methods, the information could be incorporated into the clinical trials, which might lead to substantial gain in study power and reduce the sample size needed. This talk will discuss the considerations of using alternative data sources and the common methods used to incorporate such information with case examples from rare disease and pediatric area.

Bayesian Approach of testing Noninferiority in Presence of Historical Data
Samiran Ghosh
School of Medicine, Wayne State University
sghos@med.wayne.edu
Randomized controlled trials (RCT’s) are an indispensable source of information about efficacy of treatments in almost any disease area. With the availability of multiple treatment options, comparative effectiveness research (CER) is gaining importance for better and informed health care decisions. However, design and analysis of effectiveness trial is much more complex than the efficacy trial. The effect of including an active comparator arm/s in a RCT is immense. This gives rise to superiority and non-inferiority trials. The non-inferiority (NI) RCT design plays a fundamental role in CER, which will be also focus of this talk. In the past decade many statistical methods have been developed, though largely in the Frequentist setup. However, availability of historical placebo-controlled trial is useful and if integrated in the current NI trial design, can provide better precision for CER. This may reduce sample size burden and improves statistical power significantly in current trial. Bayesian paradigm provides a natural path to integrate historical as well as current trial data via sequential learning in the NI setup. In this talk we will elucidate with example how Bayesian methods can be applied to reduce substantial sample size burden in the context of NI trial.

A Note on Posterior Predictive Assessment to Assess Model Fit for Incomplete Longitudinal Data

Arkendu Chatterjee
Merck & Co., Inc.
arkendu.chatterjee@merck.com

We examine two posterior predictive distribution based approaches to assess model fit for incomplete longitudinal data. The first approach assesses fit based on replicated complete data as advocated in Gelman et al. (2005). The second approach assesses fit based on replicated observed data. Differences between the two approaches are discussed and an analytic example is presented for illustration and understanding. Both checks are applied to data from a longitudinal clinical trial. The proposed checks can easily be implemented in standard software like (Win)BUGS/JAGS/Stan.

Key words and phrases: Missing data; Model diagnostics; Posterior predictive distribution

Session 87: Immunogenicity Testing of Therapeutic Protein Product

Issues in designing immunogenicity studies
Yun Wang
U.S. Food and Drug Administration
gracewb@gmail.com

In a biosimilar product development, in addition to analytical and nonclinical studies, comparative clinical studies may be conducted to assess whether there is clinical meaningful difference between the biosimilar products and the reference product from efficacy and immunogenicity perspective. In this presentation, the speaker will focus on some issues we encounter in designing head-to-head immunogenicity studies to evaluate the equivalence of the test product and reference product in the incidence of anti-drug antibodies (ADA). For example, in general, reference ADA rate and corresponding bioequivalence margin, derived based on reference product labeling and literature, are used in designing these studies. However, due to limited available information, the reference point estimate and the corresponding variation in the ADA rate may not be reliably estimated. Study designed using these unreasonably estimated parameters may need to be adapted later when more information is available. Another case is that for some products, due to low reference ADA rate, a reasonably small bioequivalence margin will require a large study with unfeasible sample size. In this case, we may design the study with relatively large equivalence margin and small sample size and incorporate sample size re-estimation based on interim analysis.

Statistical evaluation of several methods for cut point determination of immunogenicity screening assay
Yi Tsong1, Meiyu Shen2 and Xiaoyu Dong3

1CDER, FDA
2U.S. Food and Drug Administration
3Amgen

Cut point of the immunogenicity screening assay is the level of response of the immunogenicity screening assay at or above which a sample is defined to be positive and below which it is defined to be negative. Food and Drug Administration Guidance for Industry on Assay Development for Immunogenicity Testing of Therapeutic recommends the cut point to be an upper 95 percentile of the negative control patients. In this article, we assume that the assay data is a random sample from a normal distribution. The sample normal percentile is a point estimate with variability that decreases with the increase of sample size. Therefore the sample percentile does not assure at least 5% false positive rate with high confidence level (e.g., 90%) when the sample size is not sufficiently enough. With this concern, we propose to use a lower confidence limit for a percentile as the cut point instead. We have conducted an extensive literature review on the estimation of the statistical cut point and compare several selected methods for the immunogenicity screening assay cut point determination in terms of bias, the coverage probability, and false positive rate. The selected methods evaluated for the immunogenicity screening assay cut point determination are sample normal percentile, the exact lower confidence limit of a normal percentile (Charkraborti and Li, 2007), and the approximate lower confidence limit of a normal percentile. It is shown that the actual coverage probability for the lower confidence limit of a normal percentile using approximate normal method is much larger than required confidence level with a small number of assays conducted in practice. We recommend using the exact lower confidence limit of a normal percentile for cut point determination.

Session 88: Statistical advances and challenges in immune-oncology development

Robust Phase 3 Designs for Adaptive Population Modification
Cong Chen1, Xiaoyun (Nicole) Li, Wen Li and Robert Beckmann
1Merck & Co., Inc.
cong.chen@merck.com

It is well documented in this genomic era that an investigational new drug may have greater treatment effect in a biomarker positive population than in the biomarker negative population. However, limited by preclinical data and early phase clinical data, a lot of Phase 3 confirmatory trials are initiated without fully understanding the biomarker effect. In this presentation, we will investigate the impact of adaptive population expansion on the overall Type I error in two statistical designs. The endpoint for making the adaptive decision can be different from the primary endpoint of the study. The first design allows expansion of study population from biomarker positive patients to all-comers if the treatment effect in the biomarker positive population is more impressive than expected, suggesting broader activity of the study...
drug. We show that, under this design, the trial outcome can be tested at the desired alpha level without inflating the Type I error when the adaptive decision is based on the primary endpoint of the study or based on an endpoint non-negatively correlated with the primary endpoint, an assumption that generally holds in practice. The second design allows addition of biomarker positive patients in an all-comer study if the treatment effect in the biomarker negative population is less impressive than expected, suggesting lower probability of success in the all-comer population. We show that, under this design, the trial outcome can always be tested at the desired alpha level without inflating the Type I error.

A Bayesian hierarchical model for indirect comparison of immuno-oncology drugs

† Jingyi Liu, Ji Lin, Zach Thomas and Yumin Zhao
Eli Lilly and Company

Liu.jingyi@lilly.com

The first PD-1 checkpoint inhibitor, pembrolizumab, was approved in 2014, and subsequently the medical literature regarding cancer immunotherapy has rapidly expanded. The PD-1/PD-L1 treatment axis is characterized by: 1) the same pathway with 2 counterpart binding-sites, 2) a large number of molecules in development, 3) broad range of indications being explored, and 4) similar efficacy results within an indication. This provides a unique opportunity to assess many important clinical questions, e.g. whether antibodies targeting PD-1, the receptor located on T cells, have a different efficacy and/or safety profile from the antibodies targeting PD-L1, the ligand for PD-1. Despite the sudden increase of publications shedding light on this and other critical questions, there is only a limited focus on systematic statistical methodologies and analyses which may provide a more formal quantitative characterization. We used a meta-analytic procedure to compare efficacy and safety data between PD-1 and PD-L1 checkpoint inhibitors across tumor types. The procedure uses a Bayesian hierarchical model to synthesize efficacy and safety data between molecules across indications. The model was applied to public data collected from approximately 70 clinical studies through Dec 2016, across 31 solid tumor types and involving 12,025 patients. This yielded interesting insights including that treatment with PD-1 inhibitors resulted in slightly higher numerical response rates, but the magnitude of difference was not clinically relevant. In conclusion, it is critical to understand the ever-changing PD-1/PD-L1 landscape and optimal treatment regimens. Direct comparison of these molecules is difficult, but our proposed method provides an approach to address this critical question, and potentially others, through efficient use of publically available data.

Statistical Considerations When Applying Restricted Mean Survival Time in Immuno-Oncology Trials

† Ying Lu1, Hua Jin2 and Lu Tian1
1Department of Biomedical Data Science and Center for Innovative Study Design, Stanford University
2South China Normal University and Stanford University

ylu1@stanford.edu

One of many challenges in analyzing survival endpoint in immune-oncology RCTs is that the proportional hazards (PH) assumption is often violated. Difference in restricted mean survival time (RMST) has been proposed as a robust and interpretable alternative endpoint to hazards ratio in these trials. In this paper, we discuss three topics that are relevant to the practical use of RMST in such trials. First, we discuss the selection of restricted time window in the RMST analyses. The width of the window affects the statistical power and conclusion of trials. While the window should be determined based on clinical importance, the practical choice is often limited by the censoring time, including the study durations. Here we provide a set of statistical conditions for the upper limit of RMST, under which the relevant statistical inference based on weighted Kaplan-Meier curve is valid. We also discuss the extension of window under accelerated regression models. Second, we compare four statistical tests for the alternative hypothesis that the immune treatment group stochastically increases survival time in comparison to the control group. This is a slightly more restrictive alternative than the increasing in RMST but more general than the alternatives imposed by PH model. We can compare the stochastic ordering under the restricted time window. Third, we discuss the use of a semi-parametric regression model for adjustment of covariates in RMST comparisons in immune-oncology RCTs.

Clinical Trials and Statistical Challenges for Immuno-therapies

Yaping Wang
U.S. Food and Drug Administration

yaping.wang@fda.hhs.gov

Since Immuno-therapy and Oncology are two of the most eye-catching key words in industry, it makes immune-oncology (IO) the gold mine for all the stake holders. I would give a brief introduction of immuno-therapy and its development. After that, I will introduce some of the most commonly used statistical methods and the limitations. And then, I am going to introduce some alternative analysis methods which are proposed by industry and academia. Finally, I am going to cover some examples of future possible studies.

Session 89: Statistical learning for high-dimensional and complex data

Theory Informs Practice: Smoothing Parameters Selection for Smoothing Spline ANOVA Models in Large Samples

† Xiaoxiao Sun1, Wenxuan Zhong2 and Ping Ma2
1University of Arizona
2University of Georgia

xiaosun@uga.edu

Large samples have been generated routinely from various sources. Classic statistical models, such as smoothing spline ANOVA models, are not well equipped to analyze such large samples due to expensive computational costs. In particular, the daunting computational costs of selecting smoothing parameters render the smoothing spline ANOVA models impractical. In this talk, I will present an asymptirical (asymptotic + empirical) smoothing parameters selection approach for smoothing spline ANOVA models in large samples. The proposed method can significantly reduce computational costs of selecting smoothing parameters in high-dimensional and large-scale data. We show smoothing parameters chosen by the proposed method tend to the optimal smoothing parameters minimizing a risk function. In addition, the estimator based on the proposed smoothing parameters achieves the optimal convergence rate. Extensive simulation studies will be presented to demonstrate numerical advantages of our method over competing methods. I will further illustrate the empirical performance of the proposed approach using real data.

High-dimensional Spatially Varying Coefficient Models

† Xinyi Li, Li Wang, Dan Nettleton and Guannan Wang
1Iowa State University
2College of William & Mary

xinyi.111991@gmail.com
High-dimensional data occur very frequently and are especially common in biomedical studies including gene-environment-wide association studies. In this paper, we study the spatial heterogeneity of genetic effects on phenotype responses via high-dimensional spatially varying coefficient models (SVCMS) to explore the gene-environment interaction. We use bivariate splines to represent the coefficient functions over a two-dimensional domain. The proposed method selects the correct model with probability approaching one under regularity conditions. A natural question raised in practice is if the coefficient function is really varying over space. In this paper we present a unified approach for simultaneous sparse learning and model structure identification (i.e., varying and constant coefficients separation) for ultra-high-dimensional SVCMS where the number of covariates are possibly larger than the sample size. Our method can identify zero, nonzero constant and spatially varying components correctly and efficiently. The estimators of constant coefficients and varying coefficient functions are consistent. The performance of the method is evaluated by a few simulation examples and a real data analysis.

Differential network analysis via a novel biclustering framework

Sha Cao¹, Xiaoya Lu, Chi Zhang¹ and Na Bo
¹Indiana University School of Medicine

Complex biological processes are accomplished by the concerted functions of a set of genes, which usually show strong intercorrelations. However, such correlation patterns among the genes could be altered throughout tumor development process, as the dynamic tumor micro-environment could greatly impact the normal functions of these proteins, especially the cancer driven pathways, such as Ras/Raf signaling pathway. We consider the expression matrix on patients’ tumor tissue samples over a wide range of pathological stages for a pre-defined gene set, and detect the possible differential association patterns among the genes across cancer development. We formulate this as a bi-clustering problem, where the genes and samples are simultaneously clustered, under the framework of non-negative matrix tri-factorization. The known biological interactions among the genes, and similarities of the sample phenotypes are incorporated as graph regularization on the gene cluster matrix and sample cluster matrix. Our proposed semi-supervised biclustering method could identify key hidden structural changes for key cancer pathways.

Robust Optimal Scoring Discriminant Analysis & Discrete Wavelet Packet Transform Based Discriminant Analysis for Whole Genome Sequences

Hsin-Hsiung Huang
University of Central Florida

We develop a novel classification model which is robust against outliers by introducing a loss function to optimal scoring discriminant analysis. The discriminant vectors and scoring vectors are solved by an iteratively reweighted least squares algorithm. It inherits good properties from these two ideas for reducing the influence of these outliers on estimating the discriminant directions. In the asymptotic stability analysis, we show that the influence function is bounded and discriminant vectors and scoring vectors are both consistent as the percentage of outliers goes to zero. The experimental results are presented to confirm that the robust optimal scoring discriminant analysis is effective and efficient. In recent years, alignment-free methods have been widely applied for genome sequences comparisons, since these methods compute efficiently and provide desirable phylogenetic analysis results. These methods have been successfully combined with hierarchical clustering methods for finding phylogenetic trees. However, it may not be suitable to apply these alignment-free methods directly to the existing statistical classification methods, since there still lacks an appropriate statistical classification theory for integrating with the alignment-free representation methods. In this article, we propose a discriminant analysis method which uses discrete wavelet packet transform to represent the whole genome sequences and discriminant analysis to classify the genome sequences. We show that the proposed alignment-free representation statistics of features approximately follow normal distributions. The data analysis results indicate that the proposed method provides accurate classification in real time.

Session 101: The win ratio: what is it?

What is really the win ratio?

Gaohong Dong¹, Victoria Chang², Junshan Qiu³, Roland Matsouaka⁴ and David Hoaglin⁵
¹iStats, Inc.
²Abbvie, Inc.
³U.S. Food and Drug Administration
⁴Duke University School of Medicine
⁵University of Massachusetts Medical School

gaochangdong@istats.org

A composite endpoint combining multiple outcomes as a single endpoint is frequently used as the primary endpoint for clinical trials. Conventionally time-to-first-event of the multiple outcomes within the composite endpoint is analyzed. The win ratio (Pocock et al., 2012) considering the clinical importance order of the outcomes is appealing (e.g. death as the most important outcome is considered first, then graft loss as the next important endpoint is considered, in transplant trials). The win ratio has been applied to data analysis and study design of clinical trials as a measure of the effect of the experimental treatment relative to the control. The interpretation of the win ratio, however, is not always clear, and the way it handles ties is not the only possibility. This talk will focus on the unmatched win ratio. Using clinical trial applications, we will first demonstrate what the win ratio really is, then we connect the win ratio to the odds ratio and the hazard ratio. These are essential properties to interpret the win ratio.

A Generalized Analytic Solution to the Win Ratio

Di Li² and Gaohong Dong²
²iStats, Inc.
di.li@bms.com

With real clinical trial examples, this talk will focus on the generalized analytical solution (Dong, Li, et al., 2016) to both hypothesis testing and confidence interval construction for the win ratio. It is a generalized analytical method, which is valid for any algorithm determining wins, losses and ties. Our simulations show that the logarithm of the win ratio follows a normal distribution better. The asymptotic variance of the logarithm of the win ratio can be approximately determined analytically (Dong, Li, et al., 2016) considering the clinical importance order of the outcomes within the composite endpoint is analyzed. The win ratio (Pocock et al., 2012) considering the clinical importance order of the outcomes is appealing (e.g. death as the most important outcome is considered first, then graft loss as the next important endpoint is considered, in transplant trials). The win ratio has been applied to data analysis and study design of clinical trials as a measure of the effect of the experimental treatment relative to the control. The interpretation of the win ratio, however, is not always clear, and the way it handles ties is not the only possibility. This talk will focus on the unmatched win ratio. Using clinical trial applications, we will first demonstrate what the win ratio really is, then we connect the win ratio to the odds ratio and the hazard ratio. These are essential properties to interpret the win ratio.
Confidence interval estimations for the matched win ratio
Roland Matsouaka and Adrian Coles
1Duke University School of Medicine
2Duke University
roland.matsouaka@duke.edu

The use of composite endpoints in clinical trials has increased in recent years. The use of time-to-first event composite endpoints as analysis strategy is problematic as they tend to prioritize less severe components of the composite. As alternatives to the standard composite endpoint of time-to-first event, the win ratio (WR) and the win difference (WD) have been proposed. In this talk, we present confidence interval estimations for the matched WR and WD using two different methods. The first method is based on large-sample normal approximation to the multinomial distribution. The second combines two separate one-sample proportions confidence intervals into a hybrid interval using the method of variance estimates recovery (MOVER). We show that the latter method for constructing a confidence interval is (1) as good as the large-sample confidence interval method when the sample is large and (2) is better than both the large-sample method and the exact method when the sample is small. We demonstrate the performance of the proposed methods using simulation studies with varying sample sizes and different number of ties. Finally, we illustrate the proposed methods using data from cardiovascular disease trials.

Stratified win ratio
Junshan Qiu, Gaohong Dong, Duolao Wang and Marc Vandemeulebroeke
1U.S. Food and Drug Administration
2iStats, Inc.
junshan.qiu@fda.hhs.gov

In this talk, we will present the stratified win ratio (Dong, Qiu et al., 2017), constructed in a similar way as the Mantel-Haenszel stratified odds ratio. We will introduce a formal group of its variance estimator with a plug-in of existing or potentially new variance/covariance estimators of the number of wins for two treatment groups. We will present its statistical performance using simulation studies and real clinical trial examples. Homogeneity test to assess whether the stratum-specific win ratios are homogeneous across strata will be also discussed.

Two-Step Bayesian Multiple Classifications with Logic Expressions
Wensong Wu and Tan Li
Florida International University
wenswu@fiu.edu

In this presentation we consider a two-class classification problem, where the goal is to predict the class membership of M units based on the values of high-dimensional categorical predictor variables as well as both the values of predictor variables and the class membership of other N independent units. We focus on applying generalized linear regression models with Boolean expressions of categorical predictors. We consider a Bayesian and decision-theoretic framework, and develop a general form of Bayes multiple classification function (BMCF) with respect to a class of cost-weighted loss functions. In particular, the loss function pairs such as the proportions of false positives and false negatives, and (1-sensitivity) and (1-specificity), are considered. The best Boolean expressions are selected by a data driven procedure, where the candidates are first selected by Apriori Algorithm, an efficient algorithm for detecting association rules and frequent patterns, and the final expressions are selected by Bayesian model averaging. This two-step procedure will reduce model uncertainty in model selection and retain computational efficiency. The results will be illustrated via simulations and on a vulvovaginal candidiasis (VVC) infection diagnosis dataset.

Understanding Kernel-Embedding Based Goodness-of-Fit Tests
Krishnakumar Balasubramanian
Princeton University
kb18@princeton.edu

Reproducing kernel Hilbert space (RKHS) embedding of probability distributions has attracted considerable amount of attention in recent years in the machine learning community. It offers a general and flexible framework for hypothesis testing with complex non-Euclidean data structures. Two specific examples include (i) testing with graph/tree-valued data structures with applications to bioinformatics and (ii) testing with permutation-valued data structures with applications to analyzing ranking algorithms. Despite their popularity, fairly little is known about the statistical performance of these kernel-embedding based tests. Focussing on the goodness of fit testing problem, in this talk I will first show that a vanilla version of the kernel-embedding based test is statistically suboptimal. I will then present a novel test based on a moderated embedding and show that the new approach provides minimax optimal tests for a wide range of deviations from the null and can also be made minimax adaptive over a large collection of interpolation spaces. Joint work with Tong Li and Ming Yuan.

On the generalization and computation of data depth
Yiyuan She
Florida State University
yshe@stat.fsu.edu

The concept of data depth provides valuable tools for exploratory data analysis. This talk focuses on the generalization and computation of data depth for supervised learning in relatively high dimensions. We show that data depth is statistic driven and can be significantly extended using subspace pursuit. A new formulation of Tukey’s depth links to psi functions in robust statistics and enables us to utilize state-of-the-art optimization techniques to derive efficient algorithms with provable guarantees. Simulations and real data examples are shown to demonstrate the efficacy of the proposed methodology for inference and estimation in modern statistical applications.

Advances in chromosome conformation capture and next generation sequencing data are enabling genome-wide investigation of chromatin interactions in 3D space. Hi-C experiments generate genome-wide frequencies between pairs of loci by sequencing DNA segments ligated from loci in close spatial proximity. Two essential tasks in Hi-C modeling are to infer unobserved interaction intensity and detect significant edges in the contact graph from observed Hi-C count data. We proposed an empirical Bayes method to estimate latent interaction intensity from observed Hi-C data and then...
Session 91: Real world evidence data: applications, promises and cautions

Patient Centered Network Analytics for Integrative Mining of Heterogeneous Health Data
Fei Wang
Weill Cornell Medical College
Fei.wang2@med.cornell.edu

In the current big data era, more and more health data, such as electronic health records, medical imaging, genomics, pharmacology, environments, patient surveys, etc., are becoming readily available. Recently many research has been conducted on extracting useful clinical/biological insights from those data. However, because these data are highly heterogeneous, the existing analysis is usually just focusing on a single data source. In this talk, I will present a unified graph based perspective on analyzing different data sources. I will also demonstrate the big potential on performing integrative mining of heterogeneous health data using such a graph based framework. Case studies done from my team will be presented, followed by the discussions on challenges and future directions.

Real-World Study Design and Analysis Considerations to Control for Confounding
Hongwei Wang, Weili He and Meijing Wu
Abbvie, Inc.
hongwei.wang@abbvie.com

Real-world evidence (RWE) including prospective and retrospective observational studies plays an increasingly important role in generating insights on natural history of disease, comparative effectiveness research and supporting evidence-based decision-making by health authorities, healthcare professionals and patients. Conducting research using a broad RW database does not guarantee the generalizability to general population by itself as RW databases are different. Further, due to the nonintervention nature and lack of randomization, confounding by indication need to be carefully accounted for in group comparisons. In this talk, we reviewed some good practice to control for confounding and reduce bias at both RW study design stage and analytic phase in light of specific research questions and data sources. For the analytic framework, marginal effect model based on balance score such as propensity and conditional model of multivariate regression will be reviewed in terms of their strengths and limitations, operating characteristics and performance, tactical implementation and assessment.

Real World Evidence and Model-informed Drug Development - an antidiabetic drug cardiovascular outcome case study
Zhaoling Meng1, James Rogers2, Dimple Patel1, Qi Tang1, Jonathan Sidi2, Nadi Gaudel-Dedieu1 and David Delvart1
1Sanofi
2Metrum Research Group
zhaoling.meng@sanofi.com

Cardiovascular (CV) safety outcome study is routinely required in diabetes drug approval. Recent empagliflozin approval for CV indication provides an additional risk reduction option for type II diabetes (T2DM) patients with high CV risks [1] and, at the same time, presents a potential confounding in the CV effect assessment for future studies. Some patients might already take empagliflozin at the study start and some patients might initiate it during the study. Concomitant administration of empagliflozin with the study drug increases the CV effect assessment uncertainty, especially when there is imbalanced empagliflozin addition between treatments during the study. A CV outcome study for glucagon-like peptide-1 receptor agonists (GLP-1ra) class drug presents such a case. Although historical GLP-1ra CV outcome studies (LEADER [2] and SUSTAIN 6 [3]) can provide good assumptions for GLP-1ra CV effect compared to standard of care (SOC), currently, there is no clinical study available to assess the CV effect of concomitant administration of SGLT-2i and GLP-1ra. We analyzed real world evidence data to estimate this effect. With estimated effects, necessary models of the study drug and planned study design, clinical trial simulations (CTS) were used to assess the impact of this confounding and the study probability of success (POS) for a GLP-1ra drug CV outcome study. First, PopPK and PK/HbA1c exposure-response models were used to simulate patients’ HbA1c over time for GLP-1ra and SOC arms based on the planned CV outcome study design. Secondly, literature and internal CV outcome studies were used to understand and model antidiabetic medication addition during the study, especially imbalanced addition between treatments due to differential HbA1c control. Then, a real world claim database, Truven, was used to estimate the CV effects of SGLT-2i addition, either concomitantly to GLP1-ra or alone. SGLT-2i class was assessed in the analysis considering AstraZeneca’s CVD-REAL study [4] showing SGLT-2i, as a class, significantly reduced CV risks versus other T2DM medicines. Using the Truven database, T2DM patients started on 1st GLP-1ra were included in the analysis. Baseline characteristics matched T2DM patients never used GLP-1-ra but started other new antidiabetic medication during the same time period were included as the SOC arm. Patients’ age, gender and SGLT-2i baseline usage etc. were used in the matching. The estimated GLP-1-ra vs. SOC and GLP-1-ra + SGLT-2i vs. SGLT-2i effects were estimated and integrated in the CTS along with PK, PK/PD, concomitant medication addition models, and various design factors such as sample size, enrollment rate, event rates and study dropout rates. Based on the marketing prediction of empagliflozin patient utilization during the expected study period, different percentages of patients on empagliflozin at the study start and initiation during the study were simulated. The influential factor(s) for the study outcome were explored and identified. Imbalanced antidiabetic medication additions were consistently observed in historical GLP-1-ra CV outcome studies, where 20% and 30% concomitant antidiabetic medication were observed for GLP-1ra and SOC arms, respectively. The imbalanced additions were hypothesized as due to lack of HbA1c control in SOC arm compared to GLP-1-ra arm. An empirical concomitant antidiabetic medication addition model during the blinded study phase under differential HbA1c control of SOC and GLP-1-ra arms was established using an internal historical CV outcome study. The real world evidence data estimated a GLP-1-ra vs. SOC CV benefit 10% reduction and smaller GLP-1-ra+SGLT-2i vs. SGLT-2i CV benefit. 10% to 50% SGLT2-i patient usage prevalence were assumed and tested in the CTS. The simulation indicated a small impact of differential SGLT-2i addition during the blinded study phase unless there was a fairly large % SGLT-2i patient usage. Therefore, the percentage of patients on SGLT-2i can be monitored during the study to mitigate the risk. Conclusion: In this exercise, explicit and informative assumptions are essential in addition to appropriately established modeling framework to mimic a future CV study. Real world data was used to estimate the concomitant CV effects with/without empagliflozin and inform the CTS.
CDRH Real-World Evidence Guidance and Regulatory Experience with RWE on Medical Devices

Yuzhi (Alex) Hu

FDA

Yuzhi.Hu@fda.hhs.gov

Real-World Data (RWD) are data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources. FDA issued this RWE guidance last year to clarify how we evaluate real-world data to determine whether they are sufficient for generating the types of real-world evidence that can be used in FDA regulatory decision-making for medical devices. Increasing access to and use of real-world evidence is a part of CDRH's strategic priorities and recent legislation includes sections addressing real-world evidence as well for example the FDA Reauthorization Act (FDARA) including MDUFA IV and the 21st Century Cures Act. In this guidance, we (the FDA) firstly defined the RWE related terms and clarified the scope of application. We introduced the background and context of the publication of this guidance. Then we dived into the details of using RWE in medical device regulation, specifically on the different situations where RWE can be useful, the general considerations and caveats of using RWE, the human subjective protect prospective of it, and the general quality requirement of using RWE in a regulatory setting including data relevance and data reliability.

Due to the certain concerns, the agency did not disclose the company or product names. However, we described several real case examples of the past regulatory decisions made by leveraging RWE including labeling expansion, supporting conditions of approval, used as control group, and as supportive evidence.

CDRH believes that there is opportunity for greater use of RWD/RWE in regulatory decision making for devices. This guidance is designed to provide framework to help stakeholders assess relevance and reliability of RWE. CDRH is supporting several efforts to facilitate the development of infrastructure and tools to better access and use RWE for regulatory decision making.

Session 92: Dimension reduction and nonparametric methods

Sufficient Dimension Reduction under Dimension-reduction-based Imputation with Predictors Missing at Random

Qihua Wang1 and Xiaojie Yang

1 Academy of Mathematics and Systems Science, Chinese Academy of Sciences

qhwang@amss.ac.cn

In some practical problems, a subset of predictors are frequently subject to missingness, especially when the dimension of the predictor vector is high. For this case, the standard sufficient dimension reduction (SDR) methods cannot be applied directly to avoid the curse of dimensionality. A dimension-reduction-based imputation method is developed in this article such that any of spectral decomposition-based SDR methods for full data is applicable to the case of predictors missing at random. The sliced inverse regression (SIR) is used to illustrate this procedure. The proposed dimension-reduction-based imputation estimator of the candidate matrix for SIR, termed as DRI-SIR estimator, is asymptotically normal under some mild conditions and hence the resulting estimator of the central subspace is √n-consistent. The finite sample performance of the proposed method is evaluated through comprehensive simulations and a real data set is analyzed for illustration. It is also shown that how other two popular SDR methods, namely sliced average variance estimation (SAVE) and principal Hessian direction (PHD), are extended to the case of missing predictors with the aid of the proposed imputation procedure.

Envelope Quantile Regression

Shanshan Ding1, Zhihua Su2, Guangyu Zhu3 and Lan Wang4

1 University of Delaware
2 University of Florida
3 University of British Columbia
4 University of Minnesota

sding@udel.edu

Quantile regression offers a valuable complement of classical mean regression for robust and comprehensive data analysis in a variety of applications. We propose a novel envelope quantile regression method (EQR) that adapts a nascent technique called enveloping (Cook, Li, and Chiaromonte, 2010) to improve the efficiency of standard quantile regression. The new method aims to identify material and immaterial information in a quantile regression model and use only the material information for estimation. By excluding the immaterial part, the EQR method has the potential to substantially reduce the estimation variability. Unlike existing envelop model approaches which mainly rely on the likelihood framework, our proposed estimator is defined through a set of nonsmooth estimating equations. We facilitate the estimation via the generalized method of moments (GMM) and derive the asymptotic normality of the proposed estimator by applying empirical process techniques. Furthermore, we establish that EQR is asymptotically more efficient than (or at least as asymptotically efficient as) the standard quantile regression estimators without imposing stringent conditions. Hence, our work advances the envelope model theory to general distribution-free settings. We demonstrate the effectiveness of the proposed method via Monte-Carlo simulations and real data examples.

Asymptotics and Optimal Bandwidth Selection for Nonparametric Estimation of Density Level Sets

Wanli Qiao

George Mason University

qiawanli@hotmail.com

Bandwidth selection is crucial in the kernel estimation of density level sets. Risk based on the symmetric difference between the estimated and true level sets is usually used to measure their proximity. We provide an asymptotic Lp approximation to this risk, where p is characterized by the weight function in the risk. In particular the excess risk corresponds to an L2 type of risk, and is adopted in an optimal bandwidth selection rule for nonparametric level set estimation of d-dimensional density functions (d ≥ 1).

Transformed Variable Selection in Sufficient Dimension Reduction

Yuexiao Dong

Temple University

ydong@temple.edu

In this paper, we combine variable transformation with sufficient dimension reduction to achieve model-free variable selection. Existing model-free variable selection methods via sufficient dimension reduction requires a critical assumption that the predictor distribution is elliptically contoured. We suggest a nonparametric variable transformation method after which the predictors become normal. Variable selection is then performed based on the marginally transformed predictors. Asymptotic theory is established to support the proposed method. The desirable variable selection performance
of the proposed method is demonstrated through simulation studies and a real data analysis.

Session 93: Novel Methods to Estimating and Computing Densities for Bayesian Applications

Partition Weighted Approach for Estimating Marginal Posterior Density with Applications
*Lynn Kuo, Ya-Bo Wang, Ming-Hui Chen and Paul Lewis*
University of Connecticut
lynn.kuo@uconn.edu

The computation of marginal posterior density in Bayesian analysis is essential in that it can provide complete information about parameters of interest. Furthermore, the marginal posterior density can be used for computing Bayes factors, posterior model probabilities, and diagnostic measures. The conditional marginal density estimator (CMDE) is theoretically the best for marginal density estimation but requires the closed-form expression of the conditional posterior density, which is often not available in many applications. We develop the partition weighted marginal density estimator (PWMDE) to realize the CMDE. This unbiased estimator requires only a single MCMC output from the joint posterior distribution and the known unnormalized posterior density. The theoretical properties and various applications of the PWMDE are examined in detail. The PWMDE method is also extended to the estimation of conditional posterior densities. We carry out simulation studies to investigate the empirical performance of the PWMDE and further demonstrate the desirable features of the proposed method with two real data sets from a study of dissociative identity disorder patients and a prostate cancer study, respectively.

A Bayesian hierarchical model for related densities using Polya trees
*Li Ma and Jonathan Christensen*
Duke University
li.ma@duke.edu

Bayesian hierarchical models are used to share information between related samples and obtain more accurate estimates of sample-level parameters, common structure, and variation between samples. When the parameter of interest is the distribution or density of a continuous variable, a hierarchical model for continuous distributions is required. A number of such models have been described in the literature using extensions of the Dirichlet process and related processes, typically as a distribution on the parameters of a mixing kernel. We propose a new hierarchical model based on the Polya tree, which allows direct modeling of densities and enjoys some computational advantages over the Dirichlet process. The Polya tree also allows more flexible modeling of the variation between samples, providing more informed shrinkage and permitting posterior inference on the dispersion function, which quantifies the variation among sample densities. We also show how the model can be extended to cluster samples in situations where the observed samples are believed to have been drawn from several latent populations.

Semi-Implicit Variational Inference
*Mingyuan Zhou and Mingchang Yin*
University of Texas at Austin
mingyuan.zhou@mccombs.utexas.edu

Variational inference (VI) is an optimization based method for approximate Bayesian inference. In comparison to Markov chain Monte Carlo (MCMC), VI is often faster, easier to diagnose convergence, and more scalable to big data. However, it often considerably underestimates the variance of the posterior, making it inappropriate for applications requiring accurately estimating posterior uncertainty. In this talk, I will introduce semi-implicit VI (SIVI) that expands the commonly used analytic variational distribution family, by mixing the variational parameter with a flexible distribution. This mixing distribution can assume any density function, explicit or not, as long as independent random samples can be generated via reparameterization. Not only does SIVI expand the variational family to incorporate highly flexible variational distributions, including implicit ones that have no analytic density functions, such as those generated by propagating random noises through deep neural networks, but also sandwiches the evidence lower bound (ELBO) between a lower bound and an upper bound, and further derives an asymptotically exact surrogate ELBO that is amenable to optimization via stochastic gradient ascent. With a substantially expanded variational family and a novel optimization algorithm, SIVI is shown to closely match the accuracy of MCMC for posterior inference in a variety of Bayesian inference tasks.

Bayesian Integrative Model for Deciphering High-Dimensional Genotype-Phenotype Map
Chuanhua Julia Xing
XPrecision, LLC
julia.xing@x-precision.com

Identification of genetic risk factors based on single genes has turned out to have high false positive rate in the studies of genome-wide association. Meanwhile, it is widely acknowledged that diseases are typically triggered by the confluence of multiple genetic and external factors. We develop a Bayesian model for deciphering genotype-phenotype mapping by integrating high dimensional, correlated, and mixed data. The model relaxes the assumption of conditional independence, and the joint association can be derived between two groups of variables. The model features fast computation by optimizing the model structure and the posterior calculation, and it is applicable to categorical or continuous data. We test the proposed methods using the simulation studies. We demonstrate the advantages using the genome-wide association studies that have difficulty in deciphering complex disease-causing mechanism and genetic risk factors.

Session 94: Statistical methods for large-scale complex data

Integrative statistical causal analysis of the microbiome, metabolome and inflammation data in to understand oral cancer mechanism
*Di Wu, Siiliang Gong, Rujin Wang and Flavia Teles*
1University of North Carolina at Chapel Hill
dwu@unc.edu
2University of Pennsylvania
flavia.teles@Temple.edu

The biggest challenge posed by the increasing availability of large “omics” data in dental research and clinical fields is to systematically integrate different levels of data, as a key step toward understanding causality and regulation across these genomic and environmental levels and how they relate to disease. To integrate multilevel data from Foranconi Anemia (FA), a rare genetic disorder characterized by early and frequent occurrence of oral squamous cell carcinoma, to better comprehend the basis for oral cancer susceptibility in FA. Recent studies have demonstrated association between microbiome and carcinogenesis, thus we hypothesize that the oral microbiome may foster oral cancer development in FA. The microbial
composition using 16S rRNA gene sequencing (Illumina, MiSeq), metabolomics (MS/GC/LC, Metabolon) and inflammatory profiles (Luminex, 86 analytes) were collected from 24 pairs of siblings discordant for FA. Data dimension was reduced by selecting the significant differential abundance of bacterial species, metabolites and proteins in comparing the cases with the matched control samples. A statistical analysis pipeline was developed based on the “mediator analysis” framework to integrate the 3 levels as pathways with causal interpretation. FT, DW, supported by Fanconi Anemia Research Fund (FARF). DW, supported by NIH R01 MH101819-01.

**Pearson’s chi-squared statistics: approximation theory and beyond**

*Mengyu Xu, Danna Zhang and Weibiao Wu*

University of Central Florida

We establish a chi-square approximation theory for Pearson’s Chi-squared statistics by using a high-dimensional central limit theorem for quadratic forms of random vectors. Our high-dimensional central limit theorem or invariance principle is proved under Lyapunov-type conditions that involve a delicate interplay between the dimension, the sample size and the moment conditions. Based on our modified chi-squared statistic, we propose the concept of adjusted degrees of freedom. Our procedure is applied to goodness-of-fit test for the social life feeling data.

**Early stopping for nonparametric testing**

*Meimei Liu and Guang Cheng*

Purdue University

Early stopping of iterative algorithms is a widely used regularization method to avoid over-fitting in estimation. In this talk, we show that early stopping can also be applied to obtain the minimax optimal testing in a general nonparametric setup. Specifically, a Wald-type test statistics is developed based on iterated estimate produced by functional gradient descent algorithms in a reproducing kernel Hilbert space. A notable contribution is to develop a “sharp” data dependent stopping rule: testing optimality is obtained if and only if the number of iterations is beyond an optimal order. Such a result holds for various kernel classes, including Sobolev smoothness classes and Gaussian kernel classes.

**Left truncated mixture Gaussian distribution based modeling of single cell RNA-seq data**

*Chi Zhang and Changlin Wan*

Indiana University School of Medicine

Multiple statistical models have been developed to characterize the drop-out events and transcriptional heterogeneity of single-cell RNA-seq (scRNA-seq) data. However, most of the models focus on a better fitting of the observed data pattern rather than considering what causes the drop-out and heterogeneity from a biological perspective. Starting from the system biology of transcriptional regulation and considering the varied experimental resolution among single cells, we have recently derived that a Left Truncated Mixture Gaussian (LTMG) model can effectively reflect the characteristics of scRNA-seq data. We have tested our model on 20 published large scale scRNA-seq data and validated the LTMG model outperforms other common models. Downstream LMGT model based differential gene expression and cell type clustering methods were also developed.

**Session 95: Statistical Methods Beyond Parametrics in Biomedical Studies**

**Nonparametric Identified Methods to Handle Nonignorable Missing Data**

*Mauricio Sadinle*

University of Washington

There has recently been a lot of interest in developing approaches to handle missing data that go beyond the traditional assumptions of the missing data being missing at random and the nonresponse mechanism being ignorable. Of particular interest are approaches that have the property of being nonparametric identified, because these approaches do not impose parametric restrictions on the observed-data distribution (what we can estimate from the observed data) while allowing the estimation of a full-data distribution (what we would ideally want to estimate). When comparing inferences obtained from different nonparametric identified approaches, we can be sure that any discrepancies are the result of the different identifying assumptions imposed on the parts of the full-data distribution that cannot be estimated from the observed data, and consequently these approaches are especially useful for sensitivity analysis. In this talk I will present some recent developments in this area of research and discuss current challenges.

**Recurrent Event Analysis with Covariates Observed at Irregular, Informative Clinical Visits, with Application to Electronic Medical Records Data**

*Yifei Sun and Chiung-Yu Huang*

Columbia University

Electronic medical records (EMR) are increasingly used as a data resource for observational studies. EMR data are primarily collected for clinical encounter rather than research purpose, posing substantial challenges in secondary analysis. Our re-search is motivated by an EMR-based study on recurrent serious infections after kidney transplant. Regression methods for event risk analysis usually require the covariate processes to be completely observed throughout the follow-up period. In EMR data, the time-dependent covariates are repeatedly measured at clinical visits, and the timing of visits can substantially vary across patients and depend on an individual's disease status. Simple methods such as last covariate value carried forward could result in biased estimation, and fitting complex joint models requires additional assumptions on the covariate process. In this article, we investigate the observation mechanism of EMR, where the interplay of patients’ and physicians’ decision give rise to the observed data, and propose a novel method to fit the popular proportional rate model. We replace the unknown quantity identified by a full-data distribution that cannot be estimated from the observed data) while allowing the estimation of a full-data distribution (what we would ideally want to estimate). When comparing inferences obtained from different nonparametric identified approaches, we can be sure that any discrepancies are the result of the different identifying assumptions imposed on the parts of the full-data distribution that cannot be estimated from the observed data, and consequently these approaches are especially useful for sensitivity analysis. In this talk I will present some recent developments in this area of research and discuss current challenges.

**Some Ideas on Calibration in the Presence of Nonignorable Missing Data**

*Peisong Han*

Electronic medical records (EMR) are increasingly used as a data resource for observational studies. EMR data are primarily collected for clinical encounter rather than research purpose, posing substantial challenges in secondary analysis. Our re-search is motivated by an EMR-based study on recurrent serious infections after kidney transplant. Regression methods for event risk analysis usually require the covariate processes to be completely observed throughout the follow-up period. In EMR data, the time-dependent covariates are repeatedly measured at clinical visits, and the timing of visits can substantially vary across patients and depend on an individual's disease status. Simple methods such as last covariate value carried forward could result in biased estimation, and fitting complex joint models requires additional assumptions on the covariate process. In this article, we investigate the observation mechanism of EMR, where the interplay of patients’ and physicians’ decision give rise to the observed data, and propose a novel method to fit the popular proportional rate model. We replace the unknown quantity identified by a full-data distribution that cannot be estimated from the observed data) while allowing the estimation of a full-data distribution (what we would ideally want to estimate). When comparing inferences obtained from different nonparametric identified approaches, we can be sure that any discrepancies are the result of the different identifying assumptions imposed on the parts of the full-data distribution that cannot be estimated from the observed data, and consequently these approaches are especially useful for sensitivity analysis. In this talk I will present some recent developments in this area of research and discuss current challenges.
In missing data analysis, multiple robustness is a desirable property resulting from the calibration technique. A multiply robust estimator is consistent if any one of the multiple distribution models is correctly specified. We study how to carry out calibration to construct a multiply robust estimator when data are missing not at random (MNAR). With multiple models available, where each model consists of two components, one for data distribution for complete cases and one for missingness mechanism, our proposed estimator is consistent if any one pair of models is correctly specified.

Mediation Higher Criticism Test in Epigenomic Studies

Jincheng Shen and Xihong Lin

University of Utah

It is of increasing interest to study the underlying mechanisms of whether the effect of environmental exposures on a disease outcome is mediated through genomic and epigenomic markers, e.g., DNA methylation levels. Mediation analysis provides a powerful tool for mechanism studies, and has gained great success in social science research. There has been growing interests in employing mediation analysis in genetic and epigenetic studies to investigate causal pathways for complex disease causing processes. Here we are interested in detecting DNA methylation markers that mediate the effect of a given environmental exposure on certain health outcome. Multiple markers in the same genomic region are likely to be correlated and function in concert, thus it is desired to develop a test for the mediation effects of multiple mediators in a set. For well-defined genetic structures, it is also possible that the exposure may only affect the methylation of a sparse subset of markers, which in turn impact on the outcome of interest. However, current testing procedures may have low power in this situation. We develop the Mediation Higher Criticism (MHC) test for region based DNA methylation data to test for the mediation effects of multiple epigenetic markers by allowing for sparse and weak signals and an arbitrary correlation among the markers. The proposed test can maintain the correct Type I error both asymptotically and in simulation studies with a moderate to large number of mediators while accounting for the correlation among the mediators. It can maintain satisfactory performance under various correlation structures and signal sparsity. We further apply MHC on Normative Aging Study (NAS) and detect genomic structures whose DNA methylation levels mediates the effect of smoking on lung function.

Random Initialization and Implicit Regularization in Nonconvex Phase Retrieval

Yuxin Chen, Cong Ma, Yuejie Chi and Jianqing Fan

Princeton University

This work considers the problem of solving systems of quadratic equations, namely, recovering an object of interest from m quadratic equations / samples. This problem, also dubbed as phase retrieval, spans multiple domains including physical sciences and machine learning.

We investigate the efficiency of gradient descent (or Wirtinger flow) designed for the nonconvex least squares problem. We prove that under Gaussian designs, gradient descent - when randomly initialized - yields an accurate solution within a logarithmic number of iterations, thus achieving near-optimal computational and sample complexities at once. This provides the first global convergence guarantee concerning vanilla gradient descent for phase retrieval, without the need of (i) carefully-designed initialization, (ii) sample splitting, or (iii) sophisticated saddle-point escaping schemes. All of these are achieved by exploiting the statistical models in analyzing optimization algorithms, via a leave-one-out approach that enables the decoupling of certain statistical dependency between the gradient descent iterates and the data.

Tensor SVD: Statistical and Computational Limits

Anru Zhang

University of Wisconsin-Madison

We propose a general framework for tensor singular value decomposition (tensor SVD), which focuses on the methodology and theory for extracting the hidden low-rank structure from high-dimensional tensor data. Comprehensive results are developed on both the statistical and computational limits for tensor SVD. This problem exhibits three different phases according to the signal-to-noise ratio (SNR). In particular, with strong SNR, we show that the classical higher-order orthogonal iteration achieves the minimax optimal rate of convergence in estimation; with weak SNR, the information-theoretical lower bound implies that it is impossible to have consistent estimation in general; with moderate SNR, we show that the non-convex maximum likelihood estimation provides optimal solution, but with NP-hard computational cost; moreover, under the hardness hypothesis of hypergraphic planted clique detection, there are no polynomial-time algorithms performing consistently in general.
We investigate a simple model for social learning with two characters: a teacher and a student. The teacher's goal is to teach the student the state of the world Θ, however, the teacher herself is not certain about Θ and needs to simultaneously learn it and teach it. We examine several natural strategies the teacher may employ to make the student learn as fast as possible. Our primary technical contribution is analyzing the exact learning rates for these strategies by studying the large deviation properties of the sign of a transient random walk on ℤ.

Session 97: Recent advances in statistical analysis of high-dimensional biological data

A semi-supervised approach for predicting cell type/tissue specific functional consequences of non-coding variation using massively parallel reporter assays

Zhuai He¹, Linxi Liu, Kai Wang and Iuliana Ionita-Laza
Columbia University
statzhuai@gmail.com

Predicting the functional consequences of genetic variants is a challenging problem, especially for variants residing in non-coding regions. Projects such as ENCODE and Roadmap Epigenomics make available various epigenetic features, including histone modifications and chromatin accessibility, genome-wide in over a hundred different tissues and cell types. Meanwhile, recent developments in high-throughput assays to assess the functional impact of variants in regulatory regions (e.g. massively parallel reporter assays - MPRA, CRISPR/Cas9-mediated in situ saturating mutagenesis) can lead to the generation of high quality data on the functional effects of selected variants. We propose a semi-supervised approach, referred to as GenoNet, to jointly utilize experimentally confirmed regulatory variants (labeled variants), millions of unlabeled variants genome-wide, and more than a thousand cell type/tissue specific functional annotations on each variant to predict functional consequences of non-coding genetic variants. Through the application to several experimental datasets, we demonstrate that the proposed method significantly improves prediction accuracy compared to existing functional prediction methods, both at the organism level and at the tissue/cell type level. We further show that eQTLs and dsQTLs in specific tissues tend to be substantially more enriched among variants with high GenoNet scores, and how the GenoNet scores can be used to map regulatory variants in regions of interest, evaluate 3C interaction variants and aid in the discovery of disease associated genes through an integrative analysis of lipid phenotypes using a Metabochip dataset on 12,281 individuals.

Clustering of Functional Data

Sneha Jadhav and Shuangge Ma
Yale University
snehaj19@gmail.com

We propose a clustering method for functional data. This method uses the basis function expansion of functional data and employs penalization techniques to determine the number of clusters as well as group the data in these clusters. We perform simulations to evaluate the accuracy of our clustering method and compare it with a popular functional clustering method. We also provide an application on the Drosophila life cycle gene expression data.

STEPS: An Efficient Prospective Likelihood Approach to Genetic Association analyses of Secondary Traits in Extreme Phenotype

Guolian Kang¹, Yun Li, Matthew Smeltzer, Guimin Gao and Shengli Zhao
St. Jude Children’s Research Hospital
guolian.kang@stjude.org

It has been well acknowledged that methods for secondary trait (ST) association analyses under a case-control design (STCC) should carefully consider the sampling process to avoid biased risk estimates. A similar situation also exists in the extreme phenotype sequencing (EPS) design, which is to select subjects with extreme values of continuous primary phenotype for sequencing. EPS is commonly used in modern epidemiologic and clinical studies such as the well-known National Heart, Lung, and Blood Institute Exome Sequencing Project. Due to lack of valid statistical methods, naive generalized regression or STCC method is widely used but their validity is questionable due to difference in statistical design. Herein, we propose a general prospective likelihood framework to perform association testing for binary and continuous STs under an EPS design (STEPS), which can also incorporate covariates and interaction terms. We provide a computationally efficient and robust algorithm to obtain the maximum likelihood estimates. We also present two empirical mathematical formulas for power/sample size calculations to facilitate planning of binary/continuous STs association analyses under an EPS design. Extensive simulations and application to a genome-wide association study of Benign Ethnic Neonipenia under an EPS design demonstrate the superiority of STEPS over naive generalized regression and STCC.

Dissecting the Gene-environment Interactions via Bayesian Hierarchical models

Yu Jiang¹ and Cen Wu²
University of Memphis
¹University of Memphis
²Kansas State University
yjiang@memphis.edu

Developing statistical models for the identification of important gene-environment (GE) interactions is important in the study of complex disease. The commonly adopted marginal approaches fail to accommodate the joint effects of a large number genetic markers. The existing Bayesian variable selection approaches for joint models may have limitations in violating the "main effects, interaction" hierarchical structure and inefficient in handling high dimensional data. In this study, we propose a Bayesian sparse group lasso model using two layers of spike and slab priors to identify pivotal GE interactions and main effects. Our approach respects the strong hierarchy that when an interaction term exists, then both of the corresponding main effects will also be identified. We establish the theoretical properties of the proposed approach and demonstrate its advantage over alternatives by extensive simulations and data from Health Professionals Follow-up Study.

Session 98: Inference for large complex data

Non-asymptotic goodness-of-fit tests for Stochastic Block Models.

Vishesh Karwa
Ohio State University
vishesh.karwa@gmail.com

Stochastic Block models (SBM) with unknown block structure widely used to detect communities in real world networks. Testing the goodness-of-fit of such models is a challenging task due to...
the fact that the parameters of an SBM are usually estimated from a single observed network. Classic asymptotic tests are not valid. In this talk I will introduce three different variations of Stochastic Block Models and present a finite sample goodness-of-fit test for these models, when the block structure is unknown. The finite sample test is based on applying fisher’s exact test to SBM with known blocks. The machinery of Algebraic Statistics is used in constructing this test. In particular, a key building block for estimating the finite sample distribution of the test statistic is an algebraic tool called markov bases. I will present results of a simulation study that show the performance of the test with respect to Type I and Type II error rates.

Simple and Trustworthy Asymptotic $t$ Tests in Difference-in-Differences Regressions

Cheng Liu and Yixiao Sun
Wuhan University

The paper proposes two asymptotically valid t tests in a difference-in-differences (DD) regression when the number of time periods is large while the number of individuals can be small or large. Each of two t tests is based on a special heteroscedasticity and autocorrelation robust (HAR) variance estimator that is tailored towards the inference problem in the DD setting. The difference between the two t tests is that one is based on the sandwich variance estimator of a general form while the other is based on the sandwich variance estimator of a special form. The asymptotic distributions of both t tests depend on the smoothing parameter $K$ in the HAR variance estimator. A testing-optimal procedure for choosing $K$ for the t test with a special sandwich variance estimator is developed through minimizing the type II error subject to a constraint for the type I error of the t test. By capturing the estimation uncertainty of the HAR variance estimators, both t tests have more accurate size than the corresponding normal tests. They are also as powerful as the latter tests. Compared to the nonstandard tests that are designed to reduce the size distortion of the normal tests, the proposed t tests are as accurate but are much more convenient to use, as critical values are from the standard t table.

On Cumulative Slicing Estimation for High Dimensional Data

Cheng Wang
Shanghai Jiao Tong University

In the context of sufficient dimension reduction (SDR), sliced inverse regression (SIR) is the first and perhaps one of the most popular tools to reduce the covariate dimension for high dimensional nonlinear regressions. Despite the fact that the performance of SIR is very insensitive to the number of slices when the covariate is low or moderate dimensional, our empirical studies indicate that, the performance of SIR relies heavily upon the number of slices when the covariate is high or ultrahigh dimensional. How to select the optimal number of slices for SIR is still a longstanding problem in the SDR literature, which is a crucial issue for SIR to be effective in high and ultrahigh dimensional regressions.

In this paper, we work with an improved version of SIR, the cumulative slicing estimation (CUME) method, which does not require selecting the optimal number of slices. We provide a general framework to analyze the phase transition phenomenon for the CUME method. We show that, without sparsity assumption, CUME is consistent if and only if $p/n \to 0$, where $p$ stands for the covariate dimension and $n$ stands for the sample size. If we make certain sparsity assumptions, then the thresholding estimate for the CUME method is consistent as long as $\log(p)/n \to 0$. We demonstrate the superior performance of our proposals through extensive numerical experiments.

This talk is based on a joint work with Zhou Yu and Liping Zhu.

A feature distributed framework for large-scale sparse regression

Xiangyu Wang, Chenlei Leng and David Dunson
Google, LLC

Large-scale data with a huge number of features are increasingly encountered. An attractive approach to fit these data via regression is to divide the task into smaller ones and then combine the results. This divide and conquer strategy, however, does not work unless the correlations between features can be preserved after data partitioning. This paper presents a novel parallel computing framework named DECO for sparse linear regression to overcome this curse of correlation. By employing a simple decorrelation step in divide and conquer, our framework incurs almost minimal communication cost and thus is computationally very efficient. Our method is shown to perform similarly to the often infeasible oracle estimator in a centralized setting as if all the data were fitted on a single machine. Remarkably, this performance is achieved for elliptically distributed features including Gaussianity as a special case, for heavy tailed noises with roughly a finite second moment, for sparse and weakly sparse signals, and for most popular sparse regression methods. Remarkably, we show that an upper bound on the convergence rate of the resulting estimators does not depend on the number of machines. These claims are supported by extensive numerical studies and data analysis. Furthermore, several pieces of theoretical results are of independent interest. The code implementing DECO is freely available on github.com/wwrechard/deco.

Session 99: Data Science Intersecting with Health Policy

Composite Interaction Tree for Simultaneous Learning Optimal Individualized Treatment Rules and Subgroups

Xin Qiu and Yuanjia Wang
Columbia University

Treatment response heterogeneity has long been observed in patients affected by chronic diseases. Administering an individualized treatment rule (ITR) offers an opportunity to tailor treatment strategies according to patient-specific characteristics. Black-box machine learning methods for estimating ITRs may produce treatment rules that have optimal benefit but lack transparency and interpretability. In clinical practices, it is desired to derive a simple and interpretable ITR while maintaining certain optimality that leads to improved benefit in subgroups of patients, if not on the overall sample. In this work, we propose a tree-based robust learning method to estimate optimal piece-wise linear ITRs and identify subgroups of patients with a large benefit. We achieve these goals by simultaneously identifying qualitative and quantitative interactions through a tree model, referred to as the composite interaction tree (CITree). We show that it has improved performance comparing to existing methods on both overall sample and subgroups via extensive simulation studies. Lastly, we fit CITree to Research Evaluating the Value of Augmenting Medication with Psychotherapy (REVAMP) trial for treating major depressive disorders (MDD), where we identified both qualitative and quantitative interactions and subgroups of patients with a large benefit.
Text Mining of Policy Documents
Shin Ah Oh, Kelly Zou, Jason Pan, Ching-Ray Yu and Martin Carlsson
Pfizer Inc.

Text mining of documents and blogs can be tedious but powerful to gain insights and understandings behind an abundance of words. The goal is this research is to develop and to apply text mining methods and artificial intelligence to decipher policy documents. Several aspects of natural language processing, which are known as word stemming, frequency and cluster analysis, can be applied. These methods are illustrated on publicly-available documents on policy-related topics in the United States. Word frequencies are computed and word clouds are formed before and after an important index event historically. The involvement of policies over time is evaluated using the software, R, and other visualization tools.

The future is now, but where we should focus
Haoda Fu
Eli Lilly and Company

New technology enable alternative data types, such as unstructured data, streaming data, and big data. There are challenges and opportunities. Naively applying data science package can result in bias and misleading results which will put a risk to patients. What are the common issues? And where we should focus? This talk will provide our perspective based on recent enterprise initiative on advanced analytics and data science from Eli Lilly and Company.

Session 100: Causal Inference and Covariate Balance

Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores
Shu Yang
North Carolina State University

Causal inference with observational studies often relies on the assumptions of unconfoundedness and overlap of covariate distributions in different treatment groups. The overlap assumption is violated when some units have propensity scores close to 0 or 1, so both practical and theoretical researchers suggest dropping units with extreme estimated propensity scores. However, existing trimming methods often do not incorporate the uncertainty in this design stage and restrict inference to only the trimmed sample, due to the nonsmoothness of the trimming. We propose a smooth weighting, which approximates sample trimming and has better asymptotic properties. An advantage of our estimator is its asymptotic linearity, which ensures that the bootstrap can be used to make inference for the target population, incorporating uncertainty arising from both design and analysis stages. We extend the theory to the average treatment effect on the treated, suggesting trimming samples with estimated propensity scores close to 1.

Propensity score weighting analysis and treatment effect discovery
Liang Li
MD Anderson Cancer Center

Inverse probability weighting (IPW) can be used to estimate the average treatment effect (ATE) in propensity score analysis. When there is lack of overlap in the propensity score distributions between the treatment groups under comparison, some weights may be excessively large, causing numerical instability and bias in point and variance estimation. We study a class of modified IPW estimators that can be used to avoid this problem. These weights cause the estimand to deviate from the ATE. We provide some justification for this deviation from the perspective of treatment effect discovery. We show that when lack of overlap occurs, the modified weights can achieve substantial gains in statistical power compared with IPW and other propensity score methods. We develop analytical variance estimates that properly adjust for the sampling variability of the estimated propensity scores, and augment the modified IPW estimator with outcome models for improved efficiency, a property that resembles double robustness. Results from extensive simulations and a real data application support our conclusions. The proposed methodology is implemented in R package PSW.

Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data
Zhiqiang Tan
Rutgers University

Inverse probability weighting methods using propensity scores are widely used for estimating treatment effects from observational studies. However, such methods can perform poorly even when a propensity score model appears adequate as examined by conventional techniques. In addition, there is increasing difficulty when dealing with a large number of covariates. To address these issues, we study calibrated estimation as an alternative to maximum likelihood estimation for fitting logistic propensity score models. We show that, with possible model misspecification, minimizing the expected calibration loss involves reducing a measure of relative errors. We propose a regularized calibrated estimator by minimizing the calibration loss with a Lasso penalty. We develop a novel Fisher scoring descent algorithm and provide a high-dimensional analysis, leveraging the control of relative errors for calibrated estimation. We present a simulation study and an empirical application to demonstrate the advantages of the proposed methods compared with maximum likelihood and regularization.

Assessing covariate balance for clustered data with continuous exposures
Donna Coffman1 and Megan Schuler2
1Temple University
2RAND Corp

For clustered data structures, covariate balance could be obtained either within or across clusters depending on the research question of interest. Balancing individuals across clusters allows the estimation of the average effect. Conversely, balancing individuals on covariates within clusters allows estimation of both the average effect and the variability of this effect across clusters. Within cluster balance is only driven by individual-level covariates, as within-cluster matching guarantees balance on cluster-level covariates. In this paper, we focus on propensity score weighting for a continuous exposure in a clustered data structure. As with binary exposures in clustered data structures, balance for continuous exposures may be obtained either within or across clusters. In this paper, we focus on across cluster balance, as our primary interest is estimating the average treatment effect rather than cluster-specific effects. Using simulations, we compared several specifications of the propensity score: a random-effects model (REM), a fixed-effects model (FEM), and a single-level model. Additionally, our simulations compared the performance of marginal versus cluster-mean stabilized propensity score weights. In our results, regression specifica-

2018 ICSA Applied Statistics Symposium, New Brunswick, NJ, June 14-17 133
ions that accounted for the multilevel structure reduced bias, particularly when cluster-level confounders were omitted. Furthermore, cluster mean weights outperformed marginal weights. Our covariate balance results indicated that a key advantage of using a multilevel model (i.e., FEM or REM) for propensity score estimation is that these models account for cluster heterogeneity through the use of cluster-specific intercepts. Our simulation results examined robustness to omission of a cluster-level confounder both in the absence and presence of differential confounding and suggest that under certain conditions (e.g., when the underlying data structure is linear in the parametric sense), these models can achieve good balance with regard to cluster-level covariates, whether or not these covariates are included in the propensity score model. This robustness to omission of cluster-level confounders helps to satisfy the underlying causal inference assumption of no unmeasured confounders and may be particularly advantageous in cases in which cluster-level characteristics have not been measured or are not available to the researcher.

Session 102: Contributed Session 1

Intensity Estimation of Spatial Point Processes Based on Area-Aggregated Data
*Chi-Wei Lai and Hsin-Cheng Huang*
Institute of Statistical Science, Academia Sinica
hlai112404@hotmail.com
We consider estimation of intensity function for spatial point processes based on area-aggregated data. A standard approach for estimating the intensity function for a spatial point pattern is to use a kernel estimator. However, when data are only available in a spatially aggregated form with the numbers of events available in geographical subregions, traditional methods developed for individual-level data become infeasible. In this research, a kernel-based method will be proposed to produce a smooth intensity function based on aggregated count data. Some numerical examples will be provided to demonstrate the effectiveness of the proposed method.

Bayesian Nonparametric Policy Search with Application to Periodontal Recall Intervals
*Dipankar Bandyopadhyay*, Brian Reich, Eric Laber and Qian Guan
1Virginia Commonwealth University
bandyopd@gmail.com
Tooth loss from periodontal disease (PD) remains a major public health burden in the US. With the rising cost of dental insurance premiums, future professional dental treatment plans and checkups need to prioritize patient visits based on their individual risk scores of PD, and spend more resources monitoring and treating high-risk patients, or patients with elevated levels of PD bio-markers, such as the periodontal pocket depth (PPD). The scientific basis of the 6-months follow up/recall policy in any standard dental practice is mostly non-evidence based, and shrouded in mystery. In an attempt to de-mystify this, we consider an adaptive dynamic treatment regime (DTR) formulation for recall policy estimation under a Bayesian nonparametric (BNP) paradigm. We assume the policy for a given subject at a particular time-point be a function of the subject’s associated risk score, composed of subject-level features, such as socio-demographic and PD risk factors, and the recommended and actual visit intervals. The principled policy optimization determines the associated feature weights that maximizes a population level expected reward function (determining compliance to the recommended followup, and the effect of visit timing on the PPD), subject to a cost constraint. Assuming a flexible BNP Dirichlet process specification, our posterior simulation steps incorporating the complex optimization are implemented by Markov chain Monte Carlo sampling. Both simulation studies and application of methods to a rich database of electronic dental records from the HealthPartners HMO confirms the effectiveness of our BNP formulation over other parametric, and alternative ad-hoc procedures. To the best of our knowledge, this is the maiden formulation of an evidenced-based criteria for estimating individualized PD recall intervals, under the umbrella of personalized (or precision) dentistry.

Quadratic Discriminant Analysis by Projection
*Ruiyang Wu and Ning Hao*
University of Arizona
ruiyangwu@math.arizona.edu
Discriminant analysis, including linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), is a popular approach to classification problems. LDA and its variants have been widely used in data analysis due to its simplicity. However, it can not handle data heteroscedasticity. QDA would be an ideal tool but it is usually outperformed by LDA when the number of features is moderate or high. Dimension reduction is necessary before conducting QDA. In this talk, I will introduce a new dimension reduction and classification method based on QDA. In particular, I will discuss its properties, and compare it with other classifiers.

Statistical considerations for precision studies that evaluate qualitative assays
Changhong Song
Center for Devices and Radiological Health, FDA
changhongsong@hotmail.com
The objectives of many analytical studies are to evaluate the precision of diagnostic devices. For this type of studies, repeatability and reproducibility are usually reported to support whether the variability of a measurement is acceptable or not. This study reviews common study designs and statistical analysis methods for evaluating diagnostic devices with qualitative outputs.

Bayesian use of probability for model selection
Yi Mu
CDC
hrb3@cdc.gov
The problem of variable selection is pervasive in statistical practice. Currently, methods that select one subset of exploratory risk factors still dominate the medical research field. However, it is argued here that single model selection based on p-values (classical model selection) can give unsatisfactory results. A case-control study designs can be time- and cost-efficient, but also may falsely identify factors as significant by chance due to sampling. In this paper, we report on a simulation study designed to be similar to actual matched case-control studies and show that classical model selection is more likely to select false positive factors into the final model. For example, classical approach selected up to 8% of false positive factors while Bayesian model averaging only selected 2%. Bayesian model averaging is generally less sensitive at identifying true risk factors compared to classical model selection; however if risk factors have been defined in the literature, a situation that is very common in medical research, Bayesian model averaging can effectively use this prior information and outperform classical model selection approach. The methods are also applied and compared in the context of a matched case-control study of methicillin-resistant Staphylococcus aureus infections.
Session 103: Contributed Session 2

Prediction Interval for Autoregressive Time Series via Oracally Efficient Estimation of Multi-Step Ahead Innovation Distribution Function

Lijie Gu*, Juanjuan Kong and Lijian Yang
1 Soochow University
2 Tsinghua University

A kernel distribution estimator (KDE) is proposed for multi-step-ahead prediction error distribution of autoregressive time series, based on prediction residuals. Under general assumptions, the KDE is proved to be oracally efficient as the infeasible KDE and the empirical cumulative distribution function (cdf) based on unobserved prediction errors. Quantile estimator is obtained from the oracally efficient KDE, and prediction interval for multi-step-ahead future observation is constructed using the estimated quantiles and shown to achieve asymptotically the nominal confidence levels. Simulation examples corroborate the asymptotic theory.

On asymptotic risk of order selection in integrated autoregressive models

Shu-Hui Yu
Institute of Statistics, National University of Kaohsiung

Most order selection methods in autoregressive (AR) models are devised for stationary case. In this talk, we consider an integrated AR process whose AR order and integration order are finite but unknown, and show that AIC is asymptotically inefficient in prediction, while BIC and HQIC are asymptotically efficient. The results contrast vividly with those in Sin and Yu (2017) who consider infinite AR order. We further extend Shibata’s (1976) celebrated result on the asymptotic risk of AIC-type information to integrated AR processes, under the framework of same-realization prediction. Some simulation study shows these asymptotic results are valid for fairly small sample sizes.

Least tail-trimmed absolute deviation estimation for autoregressions with infinite/finite variance

Rongning Wu* and Yunwei Cui
1 City University of New York
2 Towson University

We propose least tail-trimmed absolute deviation estimation for autoregressive processes with infinite/finite variance. We explore the large sample properties of the resulting estimator and establish its asymptotic normality. Moreover, we study convergence rates of the estimator under different moment settings and show that it attains a super-root-n convergence rate when the innovation variance is infinite. Simulation studies are carried out to examine the finite-sample performance of the proposed method and that of relevant statistical inferences. A real example is also presented.

A Principal Stratification Approach to Evaluate the Causal Effect of A Patient Activation Intervention For Bone Health Outcomes

Yiyue Lou*, Michael Jones2, Fredric Wolinsky2, Peter Cram3 and Stephanie Edmonds4
1 Vertex Pharmaceuticals
2 University of Iowa College of Public Health
3 Toronto General Hospital Research Institute (TGHRI)
4 Woman Centered Health

The Patient Activation after DXA Result Notification (PAADRN) study is a multi-center, pragmatic randomized clinical trial that was designed to improve bone health. Participants were randomly assigned to either intervention group with usual care augmented by a tailored patient-activation Dual-energy X-ray absorptiometry (DXA) results letter accompanied by an educational brochure, or control group with usual care only. The primary analyses followed the standard ITT principle, which provided a valid estimate for the intervention assignment. However, findings might underestimate the effect of intervention because PAADRN might not have an effect if the patient did not read, remember and act on the letter. We apply principal stratification to evaluate the effectiveness of PAADRN for subgroups, defined by patient’s recall of having received a DXA result letter, which is an intermediate outcome that’s post-treatment. We perform simulation studies to compare the principal score weighting methods with the instrumental variable (IV) methods. We examine principal strata causal effects on three outcome measures regarding pharmacological treatment and bone health behaviors. Finally, we conduct sensitivity analyses to assess the effect of potential violations of relevant causal assumptions. Our work is an important addition to the primary findings based on ITT. It provides a profound understanding of why the PAADRN intervention does (or does not) work for patients with different letter recall statuses, and sheds light on the improvement of the intervention.

Change Point Detection in High Dimension

Runmin Wang and Xiaofeng Shao
University of Illinois at Urbana-Champaign

In this paper, we propose a new class of test statistics that target dense alternatives in the high dimensional setting with either one single change point or multiple change points, which seems lacking in the literature. Our test is nonparametric and requires quite mild structural assumption on the data generating process and does not impose sparsity assumption. Our test statistic is built on the basis of two recent advances: the use of U-statistic based approach in two sample mean testing in the high dimensional setting as initiated by Chen and Qin (2010), self-normalization for time series [Shao (2010, 2015), Shao and Zhang (2010)], which was mainly developed for low dimensional time series and parameters. The latter has been applied to change point detection of a low dimensional time series in Shao and Zhang (2010), but their test statistic cannot be applied to the setting p>n, and typically does not work well when p is moderately large relative to n as shown in some preliminary simulations. This motivates us to develop new test statistics that work in the high dimensional setting. Our test is tuning parameter free when the alternative is one single change point, a nice feature inherited from self-normalization; it involves one trimming parameter when the alternative has an unknown number of change points but the effect of trimming is accounted for in the limiting null distribution. To the best of our knowledge, this is the first change point test (for the mean) that targets the dense alternative in the high dimensional setting, and also the first attempt to extend the self-normalization idea [Shao (2010, 2015)] to inference for high dimensional data. Theoretically, we obtain the limiting null distribution by showing the weak convergence of a two-parameter stochastic process under sensible and mild assumptions, and further derive the power under local alternatives.
Session 104: Real-time prediction of clinical trial enrollment and event times

Blinded sample size re-estimation in trials with survival outcomes and incomplete information.
Thomas Cook
University of Wisconsin-Madison
cook@biostat.wisc.edu

In many large multicenter clinical trials, especially in cardiology, the primary outcome is a composite of fatal and non-fatal events. Study power is determined by the total number of subjects with at least one primary event and the sample size and duration of follow-up are selected to achieve the target number of events given an assumed underlying survival distribution. Furthermore, potential primary events typically require adjudication by an independent event classification committee. As the study progresses, information becomes available to assess these design assumptions, potentially allowing blinded adjustments to be made to both sample size and study duration. Unfortunately, delays in both event reporting and adjudication complicate our ability to make reliable projections. I will present estimators that allow projections under certain conditions, but will also provide examples for which these conditions fail.

Real-time prediction of event times in clinical trials - an overview
Gui-Shuang Ying
University of Pennsylvania
gsying@pennmedicine.upenn.edu

Randomized clinical trials often include planned interim analyses, in which an external monitoring committee reviews the accumulated data and determines whether it is scientifically and ethically appropriate for the study to continue. With survival-time endpoints, it is often desirable to schedule the interim analyses at the times of occurrence of specified landmark events, such as the 50th event, the 100th event, and so on. Because the timing of such events is random, and the interim analyses impose considerable logistical burdens, it can be worthwhile to predict the times of such events. With modern data management system, it has become feasible to create complete study data bases in real time, enabling us to use the accumulating data from the trial itself to make predictions about its future. This talk is to provide an overview of prediction framework for event times prediction, to review the available prediction methods based on exponential model, Weibull model and non-parametric model, and to demonstrate their use in real clinical trials.

Adaptive parametric prediction of event times in clinical trials
Lan Yu1 and Daniel Heitjan2
1Sanofi
2Southern Methodist University & University of Texas Southwestern Medical Center

Adaptive parametric prediction of event times in clinical trials: Yu Lan, Daniel F HeitjanBackground: In event-based clinical trials, it is common to conduct interim analyses at planned landmark event counts. Accurate prediction of the timing of these events can support logistical planning and the efficient allocation of resources. As the trial progresses, one may wish to use the accumulating data to refine predictions. Purpose: Available methods to predict event times include parametric cure and non-cure models and a nonparametric approach involving Bayesian bootstrap simulation. The parametric methods work well when their underlying assumptions are met, and the nonparametric method gives calibrated but inefficient predictions across a range of true models. In the early stages of a trial, when predictions have high marginal value, it is difficult to infer the form of the underlying model. We seek to develop a method that will adaptively identify the best-fitting model and use it to create robust predictions. Methods: At each prediction time, we repeat the following steps: (1) resample the data; (2) identify, from among a set of candidate models, the one with the highest posterior probability; and (3) sample from the predictive posterior of the data under the selected model. Results: A Monte Carlo study demonstrates that the adaptive method produces prediction intervals whose coverage is robust within the family of selected models. The intervals are generally wider than those produced assuming the correct model, but narrower than nonparametric prediction intervals. We demonstrate our method with applications to two completed trials: The International Chronic Granulomatous Disease study and Radiation Therapy Oncology Group trial 0129. Limitations: Intervals produced under any method can be badly calibrated when the sample size is small and unhelpfully wide when predicting the remote future. Early predictions can be inaccurate if there are changes in enrollment practices or trends in survival. Conclusions: An adaptive event-time prediction method that selects the model given the available data can give improved robustness compared to methods based on less flexible parametric models.

Session 105: Analysis of Data from Wearable Devices

Robust functional principal components analysis with application to accelerometer data
Sisheng Liu1, Chongzhi Di1 and Fang Han2
1Fred Hutchinson Cancer Research Center
2University of Washington
slliu@fredhutch.org

Accelerometers are widely used to objectively measure physical activity in biomedical studies. They collect high resolution functional data, which are often highly skewed and have outliers. To address these challenges, we propose a new robust approach for conducting functional principal components analysis (FPCA), by generalizing the recently proposed elliptical component analysis for multivariate...
data (Han and Liu, 2017). Theoretical properties of the proposed method are established. Through extensive simulation studies, the proposed robust FPCA is shown to perform well under various types of functional data. We applied the method to a large epidemiological study that recorded 7-day accelerometer data on 6500 women. Keywords: robust methods, functional data analysis, principal component analysis.

Multilevel variance components model in functional data with application in minute-level accelerometer measures

Haochang Shou
University of Pennsylvania
hshou@pennmedicine.upenn.edu

The emergence of mobile technologies, such as physical activity assessed via wearable actigraphy devices has provided an unprecedented opportunity to obtain objective evaluations of multiple physiological systems in real-time over weeks or months. However, the complexity of the devices and the high-dimensionality of the data also pose many analytic challenges to time-dependent measures. Most of the current approaches are based on summary statistics of activity that neglect the important time effects. We developed multilevel functional data analysis approaches that integrate multiple domains of complex measurements and reduce the dimensionality of the data while accounting for correlations in the repeated observations. In particular, motivated by the physical activity data observed from Brisbane adolescent twin study, we extended the traditional ACE model for a single univariate trait to functional outcomes. The method simultaneously: 1) handle various levels of correlation in the data; 2) identify interpretable traits via dimensionality reduction based on principal components; and 3) estimate relative variances that are attributed by additive genetic, shared environmental and unique environmental effects. Within-family similarities of those complex measures could also be effectively quantified.

Gait characteristics extracted from raw accelerometry data collected in free-living settings

Jacek Urbanek
Johns Hopkins School of Medicine
jurbane2@jhu.edu

Background. Wearable accelerometry devices allow collection of high-density activity data in large epidemiological studies both "in-the-lab" as well as "in-the-wild" (free-living). Such data can be used to detect and identify periods of sustained harmonic walking. This report aims to establish whether the micro- and macro-features of walking identified in the laboratory and free-living environments are associated with measures of physical function, mobility, fatigability, and fitness. Methods. Fifty-one older adults (median age 77.5) enrolled in the Developmental Epidemiologic Cohort Study in Pittsburgh, Pennsylvania were included in the analyses. The study included an "in-the-lab" component as well as 7 days of monitoring "in-the-wild". Participants were equipped with hip-worn Actigraph GT3X+ activity monitors, which collect high-density raw accelerometer data. We applied a walking identification algorithm to the data and defined features of walking, such as participant-specific walking acceleration and cadence. The association between those walking features and physical function, mobility, fatigability, and fitness was quantified using linear regression analysis. Results. Micro-scale features of walking (acceleration and cadence) estimated from "in-the-lab" and "in-the-wild" data were associated with measures of physical function, mobility, fatigability, and fitness. "In-the-lab" median walking acceleration was strongly inversely associated with physical function, mobility, fatigability and fitness. Additionally, "in-the-wild" daily walking time was inversely associated with usual- and fast-paced 400m walking time. Conclusions. The proposed accelerometry-derived walking features are significantly associated with measures of physical function, mobility, fatigability, and fitness, which provides evidence of convergent validity.

Session 106: New developments in large-scale data analysis

Sorted Concave Penalized Regression

Long Feng1 and Can-Hui Zhang
1Yale University
long.feng@yale.edu

The Lasso is biased. Concave penalized least squares estimation (PLSE) takes advantage of signal strength to reduce this bias, leading to sharper error bounds in prediction, coefficient estimation and variable selection. For prediction and estimation, the bias of the Lasso can be also reduced by taking a smaller penalty level than what selection consistency requires, but such smaller penalty level depends on the sparsity of the true coefficient vector. The sorted L1 penalized estimation (Slope) was proposed for adaptation to such smaller penalty levels. However, the advantages of concave PLSE and Slope do not subsume each other. We propose sorted concave penalized estimation to combine the advantages of concave and sorted penalizations. We prove that sorted concave penalties adaptively choose the smaller penalty level and at the same time benefits from signal strength, especially when a significant proportion of signals are stronger than the corresponding adaptively selected penalty levels. A local convex approximation, which extends the local linear and quadratic approximations to sorted concave penalties, is developed to facilitate the computation of sorted concave PLSE and proven to possess desired prediction and estimation error bounds. Our analysis of prediction and estimation errors requires the restricted eigenvalue condition on the design, not beyond, and provides selection consistency under a required minimum signal strength condition in addition. Thus, our results also sharpen existing results on concave PLSE by removing the upper sparse eigenvalue component of the sparse Riesz condition.

Deep Learning Approaches for Agricultural Plant Phenotyping Based on RGB Images and Hyperspectral Images

Zheng Xu, James Schnable and Jennifer Clarke
University of Nebraska–Lincoln
zheng.xu@unl.edu

The area of high-throughput phenotyping of crops has grown dramatically in recent years as a new way to generate larger and denser data sets to analyze complex plant biological systems. More and more phenotypes (traits) of interest can be extracted from images for downstream analysis linking phenotype, genome, environmental variables, etc. However, the broad adoption of plant phenotyping still remains a bottleneck because current approaches and software using computational and image processing techniques can only extract a limited number of standard phenotypic traits. We designed deep learning approaches with the hope to predict complex traits interesting to biologists during the plant growing period based on different types of images. RGB images were used to predict leaf number and crucial plant events during growth using standard deep learning methods. Hyperspectral images were used to predict ion concentrations (N, P, K, etc) using modified deep learning methods to favor prediction based on small samples. We compared the
performance of deep learning approaches with conventional methods in literature and deep learning approaches demonstrated better performance.

**Homogeneity tests of covariance matrices with high-dimensional longitudinal data**

*Shawn Santo¹, Ping-Shou Zhong¹ and Runze Li²*

¹Michigan State University
²Pennsylvania State University

santosh@stt.msu.edu

High-dimensional longitudinal data appear when many variables are measured repeatedly over time. A challenge with such data is the existence of both spatial and temporal dependence. Our focus is on testing the temporal homogeneity of covariance matrices in high-dimensional longitudinal data with temporospatial dependence. We propose a new test statistic and establish its asymptotic distribution. If the covariance matrices are not time homogeneous, an estimator is given for the location of the change point whose rate of convergence depends on the data dimension, sample size, and signal-to-noise ratio. Furthermore, our method extends to multiple change points via binary segmentation. We demonstrate the efficacy of our procedure through several simulation studies and a time-course microarray data set. This is joint work with Ping-Shou Zhong (Michigan State University) and Runze Li (Pennsylvania State University).

**Session 107: Machine Learning for Healthcare Applications**

**Traditional Statistics and Machine Learning using Electronic Health Records Data**

*Yirui Hu and H. Lester Kirchner*

Geisinger

yhu@geisinger.edu

The widespread adoption of electronic health records (EHR) has created a secure and reliable resource of clinical information concerning the health status of individuals. The Geisinger electronic health records (EHR) database contains medical history information of more than 1.7 million patients covering 50 counties in central PA and southern NJ. To better support healthcare decisions for the study of health conditions, interventions, and outcomes using EHR data, we implement both traditional statistical methodologies and machine learning algorithms for predictive modeling. In this talk, we will discuss healthcare applications including clustering algorithms to identify subtypes of patients referred to sleep clinic, group-based trajectory modeling to describe the heterogeneity of weight change patterns in patients with bariatric surgery, and predictive modeling to identify high-risk patients using historical EHR data.

**Clinical NLP and deep learning for disease classification and reporting in Chest X-rays**

*Yifan Peng*

National Institutes of Health

yifan.peng@nih.gov

Chest X-rays are a common radiological examination for screening and diagnosis of lung diseases. Although hospitals have accumulated a large number of radiographs and reports, it is not yet known how to effectively use them to build high precision computer-aided diagnosis systems. Recently, deep learning, a class of machine learning methods based on deep neural networks, has quickly become the state of the art in computer vision. However, they have not yet realized their full potentials on medical imaging mainly because large-scale images that are needed to train complex deep learning systems are not readily available. Manually annotating a large-scale corpus is a highly costly and time-consuming process as it requires comprehension of domain-specific medical knowledge. Therefore, there is a critical need to automatically construct a big dataset of radiographs with training labels.

To this end, we propose a novel text-mining framework to automatically generate training labels from image-associated X-ray reports. During that process, we developed a high-performance tool to distinguish negative assertions from positive ones in the reports. Unlike previous methods, a novelty of our tool lies in its utilization of various linguistic theories to boost its accuracy. It also detects uncertainty, a useful feature that is not well studied in the past. When tested on four public datasets, our tool achieved an average of 9.5% improvement in precision and 5.1% in F1-score compared to state of the art.

Using such a text-mining approach, we recently constructed and released a new dataset, ChestX-ray14, consisting of 108,948 frontal-view X-ray images with 14 text-mined disease labels. To the best of our knowledge, ChestX-ray14 is the largest publicly available medical image dataset in the scientific community.

Given such a large-scale dataset, we further proposed a deep learning model, Text-Image Embedding neural network (TieNet), to read chest X-ray images. TieNet is an innovative framework that uses both text and image representations to achieve high accuracy for chest X-ray labeling. When tested on three different datasets, TieNet significantly improved the image classification result (6% increase on average in AUC for all disease categories) compared to previous best methods.

Taken together, we expect this project will contribute to advancement in understanding of the radiological world and enhancing the clinical decision-making.

**Transformation in Real World Evidence through Machine Learning**

*Jyotsna Mehta*

founder of KEVA health

jyotsna@kevalhealth.com

Real world evidence is playing an increasingly important role in HEOR and healthcare decision-making. Data science and Digital health are seen as potential disruptors in the health care industry. In this talk, we will discuss how data science and digital data technologies are empowering organizations to transform the way real world evidence is generated and used. Topics covered will be Leveraging traditional real world evidence studies, role of data science in real world evidence development and digital health innovation in evidence generation.

**Session 108: Reasonable Possibility is Statistical Science in Drug Safety Monitoring**

**Reasonable Possibility in the Pre- vs Post-Approval Settings**

*Ed Whalen*

Pfizer Inc.

Ed.Whalen@pfizer.com

Determining the likelihood of drug relatedness to an adverse event varies from early phase work to post-approval marketing. We cover examples of sources and methods used in both the pre-approval phases and the post-approval market setting, covering a large portion of a drug’s life-cycle.
Visual Analytics Tools to Assess Possible Causality of a Drug-Induced Adverse Event of Special Interest

Krishtan Singh1, Melvin Munsaka2 and Kefei Zhou3

1 GlaxoSmithKline
2 Abbvie, Inc.
3 Theravance

Standard safety data tables and patient level data listings are often inefficient to identify important safety findings. Similarly, improper selection of graphical displays can fail to reveal important safety signals. Graphical depictions of safety data also play a big role in facilitating communication of safety results with regulators, investigators, Data Monitoring Committees, and other stakeholders concisely and more effectively than the tables. In this presentation, we will discuss basic principles of good visual analytics and how they can be used not only to summarize the safety data effectively but also help identify safety signals that could potentially be attributable to the test drug. We will demonstrate the role of visual analytics in assessing an adverse event of special interest with a focus on drug-induced liver injury.

Meta-Analysis Techniques and Safety Monitoring

Rositsa Dimova

U.S. Food and Drug Administration

Safety data are collected systematically during the pre-market drug development stage to characterize the risk profile of new drugs for benefit-risk evaluation purposes. Safety data are also collected post-approval for surveillance purposes or to study suspected associations between the use of a specific drug product and adverse events. Meta-analysis techniques are used to combine data from multiple trials/data sources during both the pre-market and post-approval stages. These techniques are especially relevant for evaluation of safety events, since most randomized controlled trials are designed to assess efficacy objectives and may not have large enough sample sizes to detect an increased safety risk. In general, depending on the type of data that the meta-analysis is applied to, it is described as aggregated or individual participant data meta-analysis. In this presentation, we will discuss some of the challenges associated with the application of meta-analysis techniques to safety data. Distinction will be made depending on whether the purpose of the meta-analysis is exploratory or confirmatory. (*This presentation reflects the views of the author and should not be construed to represent FDA’s views or policies.)
A random-effects model for multi-tissue deconvolution to estimate individual-level cell-specific gene expression

Jiebiao Wang, Bernie Devlin and Kathy Roeder

Recent research in gene expression deconvolution proposed methods for deconvolving the expression of a single tissue type into cell-type-specific expression. However, those deconvolution approaches can only provide the average expression levels over individuals. Aiming to estimate individual-level cell-specific gene expression for a large number of individuals, we develop a statistical method to deconvolve the gene expression of multiple tissue types. We assume the cell-type-specific gene expression to be random and calculate their empirical Bayes estimates through an EM (Expectation-Maximization) algorithm. The method is applied to deconvolving thirteen brain tissue types in the GTEx (Genotype-Tissue Expression) data. Simulations and data analyses demonstrate the advantages of our method over the existing deconvolution approaches for estimating the expression of a single tissue type.

Intensity normalization of MR images across subjects as a pre-processing tool to improve analysis and segmentation results of MS drawn from studying unwanted variation effects in gene expressions

Amitabha Sarkar and Russell Shinohara

Intensity normalization is a crucial pre-processing step in the analysis of any study involving MRI images of multiple modalities to draw meaningful conclusions about treatment effects for diseases using radiomic features. Commonly used techniques such as histogram matching and z scoring in the context of white matter lesions used to diagnose neurodegenerative diseases like multiple sclerosis, may potentially disrupt the pathology of the original image in order to conform to a common atlas or template diluting the signal in the data. We extend a technique called RAVEL previously used in Alzheimer’s disease to account for subject-wise random variation through control voxels to produce biologically interpretable and normalized intensities to facilitate between subject comparisons. Analogous to the acquisition of MR images, microarray data have been known to suffer from artificial variation introduced during the data collection process. Our interest is to investigate and modify these existing techniques previously used in genomics in the context of Imaging data.

Session 110: Recent advances in machine learning and causal inference

Speeding Up Latent Variable Gaussian Graphical Model Estimation via Nonconvex Optimization

Quanquan Gu, Pan Xu and Jian Ma

We study the estimation of the latent variable Gaussian graphical model (LVGGM), where the precision matrix is the superposition of a sparse matrix and a low-rank matrix. In order to speed up the estimation of the sparse plus low-rank components, we propose a sparsity constrained maximum likelihood estimator based on matrix factorization, and an efficient alternating gradient descent algorithm with hard thresholding to solve it. Our algorithm is orders of magnitude faster than the convex relaxation based methods for LVGGM.

Selective inference for effect modification

Qingyuan Zhao, Dylan Small and Ashkan Ertefaie

Effect modification occurs when the effect of the treatment variable on an outcome varies according to the level of other covariates and often has important implications in decision making. When there are hundreds of covariates, it becomes necessary to use the observed data to select a simpler model for effect modification and then make valid statistical inference. A two stage procedure is proposed to solve this problem. First, we use Robinson’s transformation to decouple the nuisance parameter from the treatment effect and propose to estimate the nuisance parameters by machine learning algorithms. Next, after plugging in the estimates of the nuisance parameters, we use the Lasso to choose a sparse model for effect modification. Compared to a full model consisting of all the covariates, the selected model is much more interpretable. Compared to the univariate subgroup analyses, the selected model greatly reduces the number of false discoveries. We show that the conditional selective inference for the selected model is asymptotically valid given the classical rate assumptions in semiparametric regression. An epidemiological application will be given to demonstrate our method.

Integrative linear discriminant analysis with guaranteed error rate improvement

Quefeng Li and Lexin Li
Multimodality data, i.e., multiple types of data collected from the same subjects, are fast emerging in many scientific areas. Numerous empirical studies have found that an integrative analysis of multimodality data often results in a better statistical performance. However, the advantages of integrative analysis have mostly been witnessed empirically, with little theoretical justification. In the context of two-class classification, we propose an integrative linear discriminant analysis method, and provide a theoretical guarantee that it enjoys a smaller classification error than running linear discriminant analysis on each individual data modality. In addition, we address the issues of outliers and missing values that frequently occur in integrative analysis. We demonstrate the efficacy of our method through simulations and a multimodality neuroimaging study of Alzheimer’s disease.

Iterative Random Forests to discover predictive and stable high-order interactions
Sumanta Basu
Cornell University
sumbose@cornell.edu
Genomics has revolutionized biology, enabling the interrogation of whole transcriptomes, genome-wide binding sites for proteins, and many other molecular processes. However, individual genomic assays measure elements that interact in vivo as components of larger molecular machines that regulate gene expression. Understanding how these high-order interactions drive gene expression presents a substantial statistical challenge. Building on Random Forests (RF), Random Intersection Trees (RIT), and through extensive, biologically inspired simulations, we developed the iterative Random Forest(s) (iRF) algorithm. iRF train a feature-weighted ensemble of decision trees to detect stable, high-order interactions with same order of computational cost as RF. We demonstrate the utility of iRF for high-order interaction discovery in two prediction problems: enhancer activity for the early Drosophila embryo and alternative splicing of primary transcripts in human derived cell lines. In Drosophila, iRF re-discovered the essential role of Zelda (Zld) in early zygotic enhancer activation, and novel third-order interactions, indicative of multi-valent nucleosomes with specific roles in splicing regulation. By decoupling the order of interactions from the computational cost of identification, iRF opens new avenues of inquiry in genome biology, automating hypothesis generation for the discovery of new molecular mechanisms from genomic data.

Session 111: Statistical Considerations of In-Vitro Bioequivalence or Adhesion Data Analysis

In vitro Permeation Testing- A scientific overview
Sarah Ibrahim
U.S. Food and Drug Administration
sibrahima@email.unc.edu
In transdermal drug delivery, drugs are intended to penetrate the human skin barrier and the effectiveness of which ideally would be assessed in a clinical setting. However, in vivo human studies are particularly complex in the early stages of drug development. For decades, alternative methods have been developed as surrogates for in vivo human skin testing. The assessment of percutaneous permeation of molecules is a key step in the evaluation of dermal or transdermal delivery systems. In vitro skin permeation testing is a critical tool for understanding rate and extent of drug permeation as well as the evaluation of delivery into the different layers of skin. IVPT models, including ex vivo human skin surgical or cadaver donors, ex vivo animal skin and artificial skin models, have been studied and numerous attempts made to standardized the protocols of testing. In this discussion, we will examine the current techniques in IVPT as well as developing in vitro-in vivo correlation and the demonstration of bioequivalence.

Application of scaled bioequivalence methods to in vitro permeation testing experiments
Kimberly Walters and Lisa Weissfeld
Statistics Collaborative, Inc.
kimberly.walters@statcollab.com
In vitro methods for bioequivalence and bioavailability can be an alternative to in vivo tests and/or well-controlled clinical trials. One such approach, in vitro permeation testing (IVPT), has been discussed in the literature and is proposed for use in industry for the submission of topical products and transdermal patches. While these methods have been applied for assessing generic products, they have not seen wide application for the use of chemistry, manufacturing, and controls (CMC) applications. This presentation describes the development of statistical methodology that was applied to test the effect of different levels of excipients in a transdermal patch. The new approach extends the methodology outlined in the literature for highly variable crossover bioequivalence experiments (Grosser et al., Tothfalusi et al.) to an IVPT experiment with multiple skin donors and multiple samples per donor.

Application of mixed models to in vitro permeation testing experiments for the assessment of dose level in transdermal patches
Lisa Weissfeld and Kimberly Walters
Statistics Collaborative, Inc.
lisa.weissfeld@comcast.net
The method of in vitro permeation testing (IVPT) in the area of chemistry, manufacturing, and controls has not been widely applied in the development of transdermal patches. As a result of this phenomenon, it is natural to ask if bioequivalence is demonstrated due to the inability of the method to discriminate small changes in the properties of a transdermal patch. This talk focuses on the use of mixed models to evaluate the results of discrimination experiments. Two different examples are discussed; the first is an experiment with patches containing three different levels of drug and the second is an experiment where the corrected release of drug is calculated over the time course of the IVPT experiment.

Session 112: Contributed Session 3

Parallelism Testing Methods and Explorations in Qualification of A Reference Standard
Jing Xiao
Merck & Co., Inc.
jjingxiao429@gmail.com
As per ICH and USP, parallelism testing are applied a lot in product comparability. Traditional method of significance test was challenged by many statisticians. Equivalence test is recommended now. Can equivalence test method work good in any case? Can we avoid visual inspection if we use equivalence test? In this talk we will...
go through a practical example of qualification of a reference standard to testify the parallelism testing results by traditional method, equivalence test method, as well as some exploration works. The conclusion is that the combination of equivalence test and exploration works can work best. This method does’t only work on qualification of a reference standard, but also can be applied to any case which need to do parallelism test of two slopes.

Evaluation for Stability Data from Large Number of Batches

Steve He1, Yijie Dong2 and James Bergum

Bristol-Myers Squibb

steve.he@bms.com

The current ICH Q1E guideline requires at least three stability batches to be tested in order to propose a retest period for a drug substance or a shelf life for a drug product. A preliminary statistical test is performed to determine whether the regression lines from different batches have a common slope. The test is conducted using a recommended significance level of 0.25 to compensate for the expected low power due to the relatively limited batch size in a typical stability study. If the stability data fails this preliminary pooling test, the estimation of expiration periods will be limited by the worst individual batch slope. However, the test and criteria applicable to three batches may not be appropriate for the scenario with more than three batches, and yet these gaps have not been addressed by the current ICH guidance. To bridge the gaps between current guidance and industry practice, we will discuss potential supplemental approaches for stability data evaluation with large numbers of batches. The first part will deal with the preliminary pooling test. The statistical power of a more comprehensive approach is adjusted to account for comparisons with more batches that will still align with the ICH stability guidance. The second part will discuss the worst profile criteria when failing the pooling test. Order statistics and other statistical techniques are implemented to recommend a more representative “conservative” batch for large number of batches. Furthermore, simulation studies demonstrate that the proposed supplemental approach stays consistent with ICH guidance.

Statistical Considerations for Analytical Method Transfer Equivalence Margin

Oluwemii Oyeniram, Qianqiu Li and Jyh-Ming Shoung

Janssen Research and Development, LLC

oyenira@its.jnj.com

Method Transfer is a part of pharmaceutical development in which an analytical procedure in one laboratory is adopted by one or more recipient laboratories. Typically, an equivalence test is performed to assess the similarity of laboratory performances in the analytical procedure. In order to assure an efficient and sustainable transfer of analytical procedures, a practically relevant and scientifically sound evaluation with corresponding acceptance criteria/eqivalence margin is crucial. The success criterion is not statistical significance, but rather analytical relevance. Statistical equivalence testing is a statistical tool for method transfer which includes both, a practical and acceptable difference between results is rather difficult. In this presentation, we will discuss the use of historical data and the application of Bayesian thinking as a guidance for the scientist/analyst in the determination of equivalence margin and provide examples to accommodate differences in designs and data structure between historical and planned studies. In the course of our discussion, we will point out some limitations, which can lead to misuse, as well as the utility of applying appropriate testing designs.

Statistical considerations in “Patient in-use” stability modeling

Monisha Dey and Robert Capen

Merck & Co., Inc.

monisha.dey@merck.com

According to EMEA guidance, “the purpose of in-use stability testing is to establish, where applicable, a period of time during which a multidose product can be used whilst retaining quality within an accepted specification once the container is opened.” This poses unique technical and practical challenges concerning modeling the data to predict what kind of stresses the product has undergone before administration. For example, data collected at room temperature conditions as part of a routine stability study, may underestimate the change the critical quality attributes experiences over time as observed during a patient in-use study. Unfortunately only limited data are often available from the latter type of study compared to the former. Also, companies will often apply the pre-approved acceptance criteria for shelf life to patient in-use stability testing, which may be problematic for critical quality attributes that trend at stressed conditions. This talk discusses considerations and approaches for jointly modeling limited patient in-use data along with data obtained from routine stability studies conducted at multiple storage conditions using linear or non-linear statistical models. A case study will be shown illustrating the challenges.Keywords: stability, patient in-use, linear and non-linear degradation modeling.

Session 113: Contributed Session 4

Gaussian copula models for spatial count data: estimation and prediction

Zifei Han1 and Victor Deoliveira2

1University of Texas at San Antonio

2Vertex Pharmaceuticals

hanzifei1@gmail.com

In this talk I will introduce the methods to conduct likelihood inference and make spatial prediction in Gaussian copula models for geospatial count data. This is a computationally challenging task because the likelihood function is only expressible as a high dimensional multivariate normal integral. An efficient sequential importance sampling method is used to evaluate both the likelihood function and the predictive density. We demonstrate the effectiveness of the algorithm via simulation studies. Alternative methods, such as the pairwise composite likelihood and a Bayesian data augmentation method are reviewed. We develop the R package gcKrige with C++ programming and parallel prediction capabilities for fast implementation of the algorithms.

ICBayes: an R package for Bayesian semiparametric regression analysis of interval-censored data

Chun Pan1, Bo Cai, Lunmeng Wang and Xiaoyan Lin

1Novartis

chunpan2003@hotmail.com

This article introduces the R package ICBayes, which aims to be a comprehensive statistical package for analyzing interval-censored data under different survival regression models. Interval-censored data occur when the time-to-event of interest is not directly observable, but falls within a time interval. The current version
of ICBayes incorporates four Bayesian functions for analyzing current status data and general interval-censored data under the proportional hazards model, the proportional odds model, and the probit model. These Bayesian functions all adopt a monotone spline for the unknown non-decreasing baseline function in each semiparametric model and are developed based on model-specific data augmentations. In addition, the package provides the log pseudo marginal likelihood for each method and thus allows users to perform model comparison and selection for a given data set. The functions are illustrated through a current status real data and a general interval-censored real data.

An R package for model fitting, model selection and the simulation for longitudinal data with dropout missingness

*Cong Xu*¹, Zheng Li², Yuan Xue¹, Lijian Zhang³ and Ming Wang⁴
1Vertex Pharmaceuticals
2Pennsylvania State University
3University of International Business and Economics
4Pennsylvania State University College of Medicine

Missing data arise frequently in clinical and epidemiological fields, in particular in longitudinal studies. This paper describes the core features of an R package wgeesel, which implements marginal model fitting (i.e., weighted generalized estimating equations, WGEE; doubly robust GEE) for longitudinal data with dropouts under the assumption of missing at random, and more importantly provides comprehensive information criteria for WGEE model selection on the mean regression and/or correlation structure. Also, this package provides a valuable tool for the simulation of longitudinal data with missing outcomes of different types (i.e., continuous, binary, count). Lastly, a real data example and extensive simulations are presented to illustrate and validate our package.

Session 114: Contributed Session 5

Distance-based and RKHS-based Dependence Metrics in High-dimension

*Changbo Zhu*¹, Shun Yao², Xiuyang Zhang³ and Xiaofeng Shao³
1University of Illinois at Urbana-Campaign
2Goldman Sachs, New York City
3Texas A & M University

In this paper, we study distance covariance, Hilbert-Schmidt covariance (aka Hilbert-Schmidt independence criterion [Gretton et al. (2008)]) and related independence tests under the high dimension, low sample size setting (HDLSS). We show that the sample distance/Hilbert-Schmidt covariance between two random vectors can be approximated by the sum of squared componentwise sample cross-covariances up to an asymptotically constant factor, which indicates that the distance/Hilbert-Schmidt covariance based test can only capture linear dependence in high dimension. As a consequence, the distance correlation based t test developed by Szekely & Rizzo (2013) for independence is shown to have trivial limiting power when the two random vectors are nonlinearly dependent but component-wisely uncorrelated. This new and surprising phenomenon, which seems to be discovered for the first time, is further confirmed in our simulation study. As a remedy, we propose tests based on an aggregation of marginal sample distance/Hilbert-Schmidt covariances and show their superior power behavior against their joint counterparts in simulations. We further extend the distance correlation based t test to those based on Hilbert-Schmidt covariance and marginal distance/Hilbert-Schmidt covariance. A novel unified approach is developed to analyze the studentized sample distance/Hilbert-Schmidt covariance as well as the studentized sample marginal distance covariance under both null and alternative hypothesis. Our theoretical and simulation results shed light on the limitation of distance/Hilbert-Schmidt covariance when used jointly in the high dimensional setting and suggest the aggregation of marginal distance/Hilbert-Schmidt covariance as a useful alternative.

Pathway and Gene Selection with Guided Regularized Random Forests

*Tyler Cook*¹, Daniel Brumley¹ and Sounak Chakraborty²
1University of Central Oklahoma
2University of Missouri

tcook14@uco.edu

Many approaches have been developed in order to model a biological outcome based on microarray data. Much focus has recently been given to incorporating gene interactions via genetic pathway information available in online databases. The additional knowledge of gene relationships may help researchers better understand the biological processes under investigation. With this project, we develop a method for pathway and gene selection based on guided regularized random forests (GRRF) that allows for the ranking of both pathways and genes in classification problems. In GRRF, variable importance scores from a random forest guide a regularization procedure to identify a subset of significant predictors. Simulation studies, as well as an analysis of a breast cancer dataset, show that our methodology is successful in identifying a compact set of important pathways and genes with a low prediction error rate.

Semiparametric Dynamic Adaptive Robust Estimations for Varying High-Dimensional Networks

*Tzu-Chun Wu and Emily Kang
University of Cincinnati
wutn@mail.uc.edu

This paper proposed Dynamic Adaptive Robust Estimators (DARE) for high dimensional observations which are associated with dynamic graphical models (DGM). We assume the networks are sparse and vary in the DGM. This method combines not only the flexibility of kernel estimation to characterize the changing network structures but also the extra robustness due to the self-tuning parameters. An efficient algorithm for both light and heavy-tailed distributions via linear programming is developed. The rates of asymptotic convergence and graphic recovery have been proved, and the performance of DGM is demonstrated through extensive simulation studies and applications to real data sets.

Gaussian Sparse Partial Membership Model with Applications to Cancer Genomics

*Wei Xia, Clint Pachayidam George and George Michailidis
University of Florida
xwnju2008@gmail.com

Cluster analyses have deep roots in statistics and computer science, as a way to organize data into a small number of homogeneous groups, thus aiding interpretation. Advancements of high throughput techniques in molecular biology yield high-dimensional, molecular profiling datasets, and pose challenges but set levels of improvements for cluster analysis. The first problem to face in such data is that disease samples may not have a crisp membership to a single cluster. The abundance of variables involved further drives us to take the classical variable selection problem. We propose a Bayesian partial membership mixture modeling framework that can
capture sample-level partial membership signatures by construction, an advance over conventional finite mixture models. We solve variable selection by a principled choice of sparse priors for clusters, which yields subspace clustering as a byproduct. However, the posterior of the model parameters is intractable, and sampling partial membership variables that lie in a probability simplex makes it difficult. We introduce an efficient Markov chain Monte Carlo scheme, combining Hamiltonian Monte Carlo and Gibbs sampling, to tackle them. We evaluated the correctness our algorithm using synthetic datasets. The results on real cancer datasets provide insights on critical oncogenes that aid classification of tumor samples into groups.

**A Randomized Algorithm for Maximum Likelihood Estimation with Spatial Autoregressive Models Applied on Social Network Data**

*Miaoqi Li and Emily Kang*

University of Cincinnati

The spatial autoregressive (SAR) model is a classical model in spatial econometrics and has become an important tool in network analysis. However, with large-scale networks, existing methods of likelihood-based inference for the SAR model becomes computationally infeasible. We here investigate maximum likelihood estimation for the SAR model with partially observed responses from large-scale networks. By taking advantage of recent developments in randomized numerical linear algebra, we derive efficient algorithms to estimate the spatial autocorrelation parameter in the SAR model. Under some regularity conditions, we show that the estimator obtained by our method, called the randomized maximum likelihood estimator (RMLE), is consistently and asymptotically normal. Compelling experimental results from extensive simulation and real data examples demonstrate empirically that RMLE outperforms the state of the art by giving smaller bias and standard error especially for large-scale problems with moderate spatial autocorrelation.

**Session 115: Contributed Session 6**

**Variable selection of lipid-environment interactions in longitudinal studies**

*Fei Zhou¹, Yu Jiang², Weiqun Wang¹ and Cen Wu¹*

¹Kansas State University
²University of Memphis

Lipid species are critical components of eukaryotic membranes. They play key roles in many biological processes such as signal transduction, cell homeostasis, and energy storage. Investigations of lipid-environment interactions, in addition to the lipid and environment main effects, have important implications in understanding the lipid metabolism and related changes in phenotype. In this study, we develop a novel penalized longitudinal variable selection method to identify important lipid-environmental interactions under the generalized estimating equation (GEE) framework. Selection of mixture of important main and interaction effects can be efficiently achieved by the proposed coordinate descent based algorithm. We have conducted extensive simulation studies to demonstrate the superior performance of our method over multiple benchmark methods, in terms of both identification accuracy and prediction performance. Analysis on high-dimensional lipid datasets from skin cancer prevention studies have led to meaningful markers that provide fresh insight into the underlying mechanism of cancer preventive effects by weight control via calorie restriction and/or exercise.

**Statistical Analysis on Clustered Mixed Recurrent Event data**

*Liang Zhu¹, Yimei Li² and Leslie Robison²*

¹UT Health
²St. Jude Children’s Research Hospital

In multi-center biomedical studies, the event of interest is often recurrent, and due to different observation process, the events can be recorded in different forms at different time, i.e., recurrent-event data with event occurrence times in some observation periods, or panel-count data with event numbers in other observation periods. Moreover, subjects from the same center might be correlated. We propose methods to analyze the mixed recurrent-event data while considering clustering effects by marginal rate models. The asymptotic distributions of the model parameters are derived, while finite-sample properties are assessed through simulation studies. The proposed methods are applied to the analysis of hospitalizations among childhood cancer survivors.

**On comparing 2 correlated C indices with censored survival data**

*Yilong Zhang¹, Xiaoxia Han² and Yongzhao Shao³*

¹Merck Research Laboratories
²Public Health Sciences Department, Henry Ford Health System
³Department of Population Health, New York University School of Medicine.

As new biomarkers and risk prediction procedures are in rapid development, it is of great interest to develop valid methods for comparing the predictive power of 2 biomarkers or risk score systems. Harrell C statistic has been routinely used as a global adequacy assessment of a risk score system, and the difference of 2 Harrell C statistics as a test statistic has been suggested in the recent literature for comparison of predictive power of 2 biomarkers for censored outcome. In this study, we showed that such a test can have severely inflated type I errors as the difference between the 2 Harrell C statistics does not converge to zero under the null hypothesis of equal predictive power measured by concordance probabilities, as illustrated by 2 counterexamples and corresponding numerical simulations. We further investigate a necessary and sufficient condition under which the difference of 2 Harrell C statistics converges to zero under the null hypothesis.

**Avoiding Bias in Internal Pilot Design for Balanced Repeated Measures**

*Xinrui Zhang¹ and Yueh-Yun Chi²*

¹Novartis
²University of Florida

Repeated measures are common in clinical trials and epidemiological studies. Designing studies with repeated measures requires reasonably accurate specifications of the variances and correlations in order to select an appropriate sample size. Underspecifying the variances leads to a sample size that is inadequate to detect a meaningful scientific difference, while overspecifying the variances results in an unnecessarily large sample size. Both lead to wastage resources and placing study participants in unwarranted risk. An internal pilot design allows sample size recalculation based on estimates of the nuisance parameters in the covariance matrix. We apply a bounding approach to ensure the maximum Type I error rate of an internal pilot design is no greater than the target, and account for the stochastic nature of the final sample size in a common class of linear mixed models. The results are useful for designing studies with repeated measures and balanced design.
Analysis of Generalized Semiparametric Varying-Coefficient Effects Models for Longitudinal Data

Li Qi\textsuperscript{1} and Yanqing Sun\textsuperscript{2}

\textsuperscript{1}Sanofi
\textsuperscript{2}University of North Carolina at Chapel Hill

The generalized semiparametric varying-coefficient effects model for longitudinal data can accommodate a variety of link functions and flexibly model different types of covariate effects, including time-constant, time-varying, and covariate-varying effects. The time-varying effects are unspecified functions of time and the covariate-varying effects are nonparametric functions of a possibly time-dependent exposure variable. We develop a semiparametric estimation procedure that uses local linear smoothing and profile weighted least squares, which requires smoothing in the two different and yet connected domains of time and the time-dependent exposure variable. The asymptotic properties of the estimators of both nonparametric and parametric effects are investigated. In addition, hypothesis testing procedures are developed to examine the covariate effects. The finite-sample properties of the proposed estimators and testing procedures are examined through simulations. The proposed methods are applied to analyze the ACTG 244 clinical trial to investigate the effects of antiretroviral treatment switching in HIV infected patients before and after developing the T215Y antiretroviral drug resistance mutation.
Index of Authors

Airoldi, E. 39, 62
Allen, N. 41, 68
Alpert, K. 41, 70
Amos, C. 54, 118
An, L. 59, 63
Anderson, K. 48, 95
Afanakis, K. 41, 76

Badsha, M. 42, 71
Bai, X. 40, 45, 66, 67
Bailey-Wilson, J. 55, 118
Balachandar, R. 41, 76
Balasubramanian, K. 56, 125

Bandyopadhyay, D. 58, 134
Bao, L. 40, 46, 66, 67
Barnett, J. 39, 64
Basickes, S. 54, 78
Basu, S. 60, 41
Beckman, K. 55, 122
Beg, M. 41, 70
Bellec, P. 42, 73
Benedetti, M. 41, 70
Bennett, D. 41, 70
Bennett, R. 47, 90
Bergum, J. 60, 142
Bhalia, M. 41, 70
Biard, L. 45, 83
Bible, J. 43, 76
Bien, J. 53, 113
Blaser, M. 47, 89
Bo, N. 55, 124
Bowman, D. 60, 142
Braun, D. 53, 116
Bruce, L. 51, 106
Bruce, S. 53, 14, 118
Brumley, D. 60, 143
Buyse, D. 53, 114

Cai, B. 60, 142
Cai, G. 47, 92
Cai, J. 46, 87
Cai, L. 49, 100
Cai, T. 52, 54, 111, 119
Cao, J. 51, 108
Cao, S. 55, 124
Cao, X. 60, 139
Capen, R. 60, 142
Carlsson, M. 58, 133

Carrion, C. 51, 108
Carroll, R. 51, 106
Caswell, C. 47, 93
Chakraborty, S. 60, 143
Chan, D. 51, 108
Chan, K. 52, 111
Chang, J. 58, 110
Chang, V. 55, 124
Chatterjee, A. 55, 122
Chen, A. 51, 53, 108, 113
Chen, B. 51, 108
Chen, C. 52, 53, 55, 12, 114
Chen, F. 46, 87
Chen, H. 52, 111
Chen, K. 54, 47, 62, 89
Chen, M. 47, 49, 51, 57, 92
Chen, Q. 49, 99
Chen, S. 44, 79
Chen, W. 53, 114
Chen, X. 46, 46, 55, 88, 89
Chen, Y. 43, 48, 51, 57, 75
S6, 106, 130
Chen, Z. 49, 97
Cheng, B. 48, 53
Cheng, G. 55, 57, 120, 129
Cheng, J. 52, 52, 110, 110
Cheng, Y. 46, 86
Chi, Y. 57, 61, 130, 144
Chien, L. 40, 65
Chinchilli, V. 44, 81
Chiu, C. 54, 54, 118, 118
Choi, D. 50, 102
Choirat, C. 53, 116
Chowdhury, M. 41, 68
Christensen, J. 57, 128
Christian, B. 41, 71
Clarke, J. 59, 137
Coffman, D. 58, 133
Cohen, A. 41, 71
Coles, A. 56, 123
Conlon, E. 50, 104
Cook, D. 44, 79
Cook, R. 52, 52, 53, 109
109, 113
Cook, T. 59, 60, 136, 143
Cooner, F. 74

Coyle, P. 43, 76
Cram, P. 58, 135
Cui, S. 43, 77
Cui, Y. 42, 58, 74, 135, 139
Cui, Z. 44, 82
Cunanan, K. 40, 65

Dai, X. 51, 53, 106, 114
Dang, Q. 50, 103
Dasgupta, S. 46, 86
Datta, G. 51, 107
Datta, S. 42, 73
Deane, P. 47, 90
Devi, A. 51, 108
Delbert, D. 56, 126
Dempsey, W. 43, 75
Deng, Q. 52, 40, 47, 66, 67
Deoliveira, V. 60, 142
Devlin, B. 60, 140
Dey, M. 60, 142
Di, C. 48, 13, 59, 75, 78, 136
Di, Q. 53, 116
Dimora, V. 59, 139
Ding, J. 40, 67
Ding, S. 44, 56, 79, 127
Ding, Y. 55, 114
Dominici, F. 53, 116
Dong, G. 56, 56, 124, 124
125
Dong, H. 55, 120
Dong, X. 55, 122
Dong, Y. 56, 60, 127, 142
Du, W. 49, 97
Du, Y. 46, 59, 88, 139
Duan, R. 51, 106
Dunson, D. 58, 132
Dutta, S. 51, 107

Eaton, J. 46, 87
Edlefsen, P. 49, 98
Edmonds, S. 55, 135
Egleston, B. 44, 78
Eftekhari, A. 60, 140
Fan, J. 45, 46, 58, 73, 88
130
Fan, R. 54, 54, 118, 118
Fang, E. 54, 119

Fang, H. 54, 118
Fang, Y. 40, 47, 66, 90
Faries, D. 44, 82
Fearnhead, P. 50, 51, 101
105
Feng, L. 59, 137
Fleming, T. 46, 86
Fonnesbeck, C. 49, 99
Fu, A. 42, 71
Fu, H. 42, 74, 74, 91
Gadkari, A. 49, 97
Gajewski, B. 59, 136
Gallo, P. 44, 72, 92
Ganju, J. 40, 69
Gao, F. 55, 121
Gao, G. 57, 131
Gascier-Markolf, B. 48

Gaudel-Dedieu, N. 56, 126
Gause, C. 48, 95
Gayanano, I. 43, 76
Geng, J. 40, 67
George, C. 56, 134
George, E. 60, 139
Ghosh, S. 55, 121
Gonen, M. 50, 65
Gong, S. 57, 128
Gou, J. 48, 79
Gregg, M. 42, 73
Gu, L. 58, 135
Gu, Q. 60, 140
Gu, X. 41, 69
Gu, Y. 59, 63
Guan, Q. 58, 134
Guan, Y. 41, 69
Guo, H. 47, 90
Guo, Z. 52, 54, 111, 119

Ha, I. 50, 102
Hall, M. 53, 114
Han, D. 45, 83
Han, F. 44, 59, 80, 136
Han, P. 57, 129
Han, X. 54, 61, 117, 144
Han, Z. 60, 142
Handen, B. 41, 70
Hanko, V. 41, 70
Hannig, J. 39, 51, 98, 106

See you in Raleigh, 
North Carolina in 2019

www.icsa.org