2019 ICSA
APPLIED STATISTICS
SYMPOSIUM

Raleigh, NC
June 9-12, 2019

International Chinese Statistical Association
International Chinese Statistical Association

Applied Statistics Symposium

2019

CONFERENCE INFORMATION, PROGRAM AND ABSTRACTS

June 9 - 12, 2019
The Raleigh Convention Center
Raleigh, NC, USA

Organized by
International Chinese Statistical Association
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Welcome</td>
<td>1</td>
</tr>
<tr>
<td>Conference Information</td>
<td>2</td>
</tr>
<tr>
<td>ICSA Officers and Committees</td>
<td>2</td>
</tr>
<tr>
<td>Conference Committees</td>
<td>5</td>
</tr>
<tr>
<td>Student Volunteers</td>
<td>7</td>
</tr>
<tr>
<td>Transportation and Parking</td>
<td>8</td>
</tr>
<tr>
<td>Map</td>
<td>9</td>
</tr>
<tr>
<td>Conference Venue Information</td>
<td>10</td>
</tr>
<tr>
<td>Dining and WiFi Information</td>
<td>12</td>
</tr>
<tr>
<td>Social Events</td>
<td>13</td>
</tr>
<tr>
<td>Program Overview</td>
<td>14</td>
</tr>
<tr>
<td>Keynote Lectures</td>
<td>15</td>
</tr>
<tr>
<td>Banquet Speaker</td>
<td>18</td>
</tr>
<tr>
<td>Student Paper Awards</td>
<td>19</td>
</tr>
<tr>
<td>Short Courses</td>
<td>20</td>
</tr>
<tr>
<td>Sponsors and Career Services</td>
<td>27</td>
</tr>
<tr>
<td>Announcements</td>
<td>38</td>
</tr>
<tr>
<td>ICSA 2020 Applied Statistics Symposium</td>
<td>38</td>
</tr>
<tr>
<td>ICSA China 2020</td>
<td>39</td>
</tr>
<tr>
<td>Scientific Program</td>
<td></td>
</tr>
<tr>
<td>June 10 9:10-10:10</td>
<td>40</td>
</tr>
<tr>
<td>June 10 10:30 - 12:10</td>
<td>40</td>
</tr>
<tr>
<td>June 10 14:00 - 15:40</td>
<td>42</td>
</tr>
<tr>
<td>June 10 16:00 - 17:40</td>
<td>45</td>
</tr>
<tr>
<td>June 11 8:00-9:00</td>
<td>47</td>
</tr>
<tr>
<td>June 11 9:10-10:10</td>
<td>48</td>
</tr>
<tr>
<td>June 11 10:30 - 12:10</td>
<td>48</td>
</tr>
<tr>
<td>June 11 14:00 - 15:40</td>
<td>50</td>
</tr>
<tr>
<td>June 11 16:00 - 17:40</td>
<td>53</td>
</tr>
<tr>
<td>June 12 08:30 - 10:10</td>
<td>56</td>
</tr>
<tr>
<td>June 12 10:30 - 12:10</td>
<td>59</td>
</tr>
<tr>
<td>Abstracts</td>
<td></td>
</tr>
<tr>
<td>Keynote Lecture I</td>
<td>62</td>
</tr>
<tr>
<td>Session 1: Modern Methods for Neuroimaging Data</td>
<td>62</td>
</tr>
<tr>
<td>Session 2: Statistical Advances in Genetics, Genomics and Bioinformatics</td>
<td>63</td>
</tr>
<tr>
<td>Session 3: Recent development in missing data and survey sampling problems</td>
<td>64</td>
</tr>
<tr>
<td>Session 4: Advances in statistical methods for large-scale genomic studies</td>
<td>65</td>
</tr>
<tr>
<td>Session 5: Analyzing big and complex data using modern machine learning techniques</td>
<td>66</td>
</tr>
<tr>
<td>Session 6: Recent advances in statistical network modeling: go beyond simple network data</td>
<td>67</td>
</tr>
<tr>
<td>Session 7: Recent Advances in Empirical Likelihood Methods</td>
<td>67</td>
</tr>
<tr>
<td>Session 8: Functional data analysis: Foundations, Tools and Applications</td>
<td>68</td>
</tr>
<tr>
<td>Session 9: Innovative methods for complex survival data and the applications</td>
<td>68</td>
</tr>
<tr>
<td>Session 10: Classification and modeling of multiple time series and functional data and applications</td>
<td>69</td>
</tr>
<tr>
<td>Session 11: Statistical Advances in High Dimensional Data Analysis</td>
<td>70</td>
</tr>
<tr>
<td>Session 12: Challenges in implementation of ICH E17</td>
<td>71</td>
</tr>
</tbody>
</table>

The 2019 ICSA Applied Statistics Symposium is held from Sunday June 9 to Wednesday June 12, 2019 at the Raleigh Convention Center, Raleigh, North Carolina, USA. This is the 28th annual symposium for ICSA. The theme of this conference is *Modernizing Statistics for Emerging Data Science*, in recognition of a new data science era for statisticians with rising opportunities and challenges.

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 9 short courses and 103 scientific sessions, including three keynote lectures, one featured session and two student paper award sessions, as well as exciting social events. Keynote lectures are from three distinguished statisticians from government, industry, and academia: Dr. Lilly Yue (U.S. Food and Drug Administration), Dr. Stephen Ruberg (Analytix Thinking), and Dr. Marie Davidian (North Carolina State University). The symposium highlights methodological and applied contributions of statistics, mathematics, and computer sciences. It brings together the statistical community and scientists from related fields to present, discuss and disseminate research and best practical practice.

With your full support, this symposium attracts more than 400 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 2019 ICSA Applied Statistics Symposium include the opening mixer (Sunday, June 9 evening) and banquet (Tuesday, June 11 evening) with the banquet speak from a renowned scholar and statistician, Dr. David Banks (Duke University). We believe this conference will be a memorable, interesting and enjoyable experience for all of you.

Raleigh is the capital and second largest city of North Carolina, and 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. Downtown Raleigh provides numerous opportunities for dining, shopping and lodging, etc. Finally, the conference hotel Marriott Raleigh City Center is next to the Raleigh Convention Center. It is our sincere hope you take the opportunity to experience these wonderful activities during your stay in Raleigh.

**Thank you for coming to the 2019 ICSA Applied Statistics Symposium in Raleigh!**

Wenbin Lu, on behalf of 2019 Applied Statistics Symposium Executive and Organizing Committees
ICSA Officers and Committees

ICSA 2019 EXECUTIVES AND MEMBERS OF THE COMMITTEES

EXECUTIVES
President: Heping Zhang (heping.zhang@yale.edu)
Past President: Aiyi Liu (aiviliu65@gmail.com)
President-elect: Jianguo (Tony) Sun (sunji@missouri.edu)
Executive Director: Gang Li (2017-2019, gli@its.jnj.com)
ICSA Treasurer: Rochelle Fu (2019-2021, hongliangshi15@gmail.com)
The ICSA Office Manager: Grace Ying Li,
Email: oicsa@icsa.org, Phone: (317) 287-4261.

BOARD of DIRECTORS
Jie Chen (2017-2019, jiechen0713@gmail.com),
Jianhua Hu (2017-2019, jh3992@cuncolumbia.edu),
Bo Huang (2017-2019, bo.huang@pfizer.com),
Chuang-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.lieh@gmail.com),
Mark Chunming Li (2018-2020, Chunming.M.Li@pfizer.com),
Niansheng Tang (2018-2020, nstang@yvu.edu.cn),
Yinglei Lai (2019-2021, ylai@gwu.edu),
Lei Shen (2019-2021, shen_lei@lilly.com),
Yifei Sun (2019-2021, yuks3072@cuncolumbia.edu),
Xin Tian (2019-2021, xin@nhlbi.nih.gov),
Kelly Zou (2019-2021, KelZouDS@gmail.com)

Biometrics Section Representative
Mengling Liu (2019, Mengling.Liu@nyumc.org)

STANDING COMMITTEES

Program Committee
Peter Song (Chair, 2019, pxsong@umich.edu), Wenguang Sun (2017-2019, USC, wenguans@marshall.usc.edu), Hong Tian (2017-2019, HTian@its.jnj.com); Mengling Liu (2018-2020, Mengling.Liu@nyumc.org), Xiwu Lin (2018-2020, xlin38@ITS.JNJ.com), Wenbin Lu (2018-2020, wlu4@ncsu.edu); Guoqing Diao (2018-2020, gdiao@emu.edu), Kai Yu (2018-2020, yuka@mail.nih.gov), Liuquan Sun (2018-2020, slq@amt.ac.cn), Ji Zhu (2019-2021, jizhu@umich.edu), Bin Nan (2019-2021, nanb@uci.edu), Alan Y Chiang (2019-2021, achiang@celgene.com).

Finance Committee
Rochelle Fu (Chair, 2019-2021, fur@ohsu.edu), Hongliang Shi (hongliangshi15@gmail.com), Rui Feng (ruifeng@pennmedicine.upenn.edu), Shu Yang (syang24@ncsu.edu).
**Nomination for Election Committee**

Annie Qu (Chair, 2019, anniequ@illinois.edu); Hongzhe Li (2017-2019, hongzhe@mail.med.upenn.edu); Niansheng Tang (2017-2019, nstawt@yale.edu); Chin-Tsang Chiang (2018-2020, chiangct.math.ntu.edu.tw); Jeen Liu (2018-2020, jiemyichuan@yahoo.com.cn); Xiao Tian (2018-2020, tianx@nhlbi.nih.gov); Ming Tan (2018-2020, Ming.Tan@georgetown.edu); Joayang Sun (2018-2020, jsun@case.edu); Chunling Liu (2019-2021, catherine.chunling.liu@polyu.edu.hk); Yiyuan She (2019-2021, yshe@stat.fsu.edu); Bo Huang (2019-2021, Bo.Huang@pfizer.com).

**Publication Committee**

Yichuan Zhao (2019, yichuan@gsu.edu), Yi Huang (Editor of Bulletin, yihuahuang@umbc.edu), Mei-Cheng Wang (Editor of SIB, mcwang@jhu.edu), Hongzhe Li (Editor of SIB, hongzhe@mail.med.upenn.edu), Hsien-Cheng Huang, (Editor of S. Sinica), Ruy Tsay (Editor of S. Sinica), Zhiliang Ying (Editor of Bulletin), Jiahua Chen (Editor of ICSA book series), Din Chen (Co-Editor of ICSA book series).

**Membership Committee**

Bo Fu (Chair, 2019, bo.fu@abbvie.com), Pengsheng Ji (2017-2019, psji@uga.edu), Yong Chen (2017-2019, ychen123@upenn.edu), Rui Feng (2017-2019, rufeng@mail.med.upenn.edu), Mark Li (2018-2020, Chunming.M.Li@pfizer.com), Yaping Wang (2018-2020, vapping.wang@fda.hhs.gov), Chenguang Wang (2018-2020, cwang68@jhmi.edu), Niansheng Tang (2019-2021, nstawt@yale.edu.cn), Lei Shen (2019-2021, shen lei@lilly.com).

**Special Lecture Committee**

Jianqiong Fan (Chair, 2019, jiangqiong.fan@outlook.com), Qiman Shao (2017-2019, CUHK, qms@sta.cuhk.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@columbia.edu), Gang Li of UCLA, (2018-2020, vli@g.ucla.edu), Haiqun Lin (2019-2021, hqlin.nyu.edu), Huazhen Lin (2019-2021, linhz@swufe.edu.cn).

**Awards Committee**

Yufeng Liu (Chair, 2019, University North Carolina, yfliu@email.unc.edu), 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, huibinzhou@yale.edu), Frank Guanghan Liu (2018-2020, guanghan.frank.liu@merck.com), Song Yang (2018-2020, yangso@nhlbi.nih.gov), Mei-Cheng Wang (2018-2020, mcowang@jhmi.edu), Judy Wang (2019-2021, jdywangle@gwu.edu), Ning Hao (2019-2021, nhao@math.arizona.edu), Jie Chen (2019-2021, jiechen0713@gmail.com).

**Financial Advisory Committee**

Rui Feng (Chair, rufeng@upenn.edu), Hongliang Shi (hongliangshi15@gmail.com), Nianjun Liu, (liunjian@indiana.edu), Xiangqin Cui (xiangqin.cui@emory.edu), Xiangqin Cui (xiangqin.cui@emory.edu), Fang Chen (FangK.Chen@sas.com), Rochelle Fu (fur@ohsu.edu).

**IT Committee**

Chengsheng Jiang (Chair, 2019, chengsheng.jiang@gmail.com), Jessica Jin (jinxu@hotmail.com), Don Sun (dsun48@gmail.com), Hongtu Zhu (htzhu@email.unc.edu).

**Archive Committee**

Chung-Chou H (Joyce) Chang (Chair, 2019, changji@pitt.edu), Xin Tian (2019-2021, tianx@nhlbi.nih.gov), Grace Ying Li (oicsa@icsa.org).

**Lingzi Lu Award Committee**

Chan, Ivan (2019-2021, ivan.chan@abbvie.com), Jichun Xie (2018-2020, Duke University, jichun.xie@duke.edu), Hao Chen (2017-2019, UC Davis, hxchen@ucdavis.edu), Victoria Sides (2017-2019, victo riasodes16@gmail.com), Shelly Hurwitz (2017-2019, hurwitz@hms.harvard.edu), Eric Kolaczyk (2014-2017, kolaczyk@math.bu.edu).
ICSA Officers and Committees

ICSA Representative to JSM Program Committee
Zhen Chen (2019, chenzhe@mail.nih.gov), Xuming He (2020, xmhe@umich.edu).

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

AD HOC COMMITTEES

2019 Applied Statistics Symposium
Wenbin Lu, (mxie@stat.rutgers.edu) and Mengling Liu (Mengling.Liu@nyumc.org) Co-Chairs of the Executive Committee

2019 ICSA China Conference
Zhezhen Jin (zj7@cumc.columbia.edu), Chair of the Program Committee.

2019 ICSA International Conference
Hongzhe Li (hongzhe@mail.med.upenn.edu), Chair of the Program Committee.

2019 JSM Local Committee
Fuyong Xing and Fan Yang (edmundxing@gmail.com and fan.3.yang@ucdenver.edu, Colorado University)

CHAPTERS

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

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

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

Executive Committee

- 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

Scientific Program Committee

- Haoda Fu, Eli Lilly and Company
- Shuyen Ho, UCB Biosciences Inc.
- Jessie Jeng, North Carolina State University
- Chunming (Mark) Li, Pfizer
- Mengling Liu, New York University
- Wenbin Lu, North Carolina State University
- Timothy O’Brien, Loyola University Chicago
- Xiaohua Sheng, FMD K & L
- Rui Song, North Carolina State University
- Yang Song, Vertex Pharmaceuticals Inc.
- Jung-Ying Tzeng, North Carolina State University
- Luo Xiao, North Carolina State University
- Yunling Xu, FDA
- Yonggang Yao, SAS Inc.
- Qing Yang, Duke University
- Shu Yang, North Carolina State University
- Donglin Zeng, University of North Carolina at Chapel Hill
- Bo Zhang, IBM
- Daowen Zhang, North Carolina State University
- Jiajia Zhang, University of South Carolina
- Kai Zhang, University of North Carolina at Chapel Hill
- Lanju Zhang, AbbVie Inc.
- Yichuan Zhao, Georgia State University
Student Paper Competition Committee

- Luo Xiao, Chair, North Carolina State University
- Gen Li, Columbia University
- Meng Li, Rice University
- Suyu Liu, University of Texas MD Anderson Cancer Center
- Min Qian, Columbia University
- Rui Song, North Carolina State University
- Yang Song, Vertex Pharmaceuticals Inc.
- Yanxun Xu, Johns Hopkins University
- Lingzhou Xue, Pennsylvania State University
- Shu Yang, North Carolina State University
- Jiajia Zhang, University of South Carolina
- Yichuan Zhao, Georgia State University
- Yingqi Zhao, Fred Hutchinson Cancer Research Center
- Ruoqing Zhu, University of Illinois at Urbana-Champaign
- Yunzhang Zhu, Ohio State University

Short Course Committee

- Yonggang Yao, Chair, SAS Inc.
- Jung-Ying Tzeng, North Carolina State University
- Lanju Zhang, AbbVie Inc.

Funding Raiser Committee

- Haoda Fu, Chair, Eli Lilly and Company
- Chunming (Mark) Li, Pfizer
- Xiaohua Sheng, FMD K & L

Local Committee

- Qing Yang, Chair, Duke University
- Bo Zhang, IBM
- Xinming An, University of North Carolina at Chapel Hill

Website

- Chengsheng Jiang, University of Maryland, College Park
- Hengrui Cai, North Carolina State University
The organizing committee appreciates the time and great help from wonderful local student volunteers. In particular, this program book is made through tremendous efforts from Duyeol Lee (UNC Chapel Hill) and Hang Yu (UNC Chapel Hill).

Below is the complete list of student volunteers for this conference:

- Sukanya Bhattacharyya (NCSU)
- Joyce Cahoon (NCSU)
- Meichen Dong (UNC Chapel Hill)
- Enying Gao (UNC Chapel Hill)
- Beilin Jia (UNC Chapel Hill)
- Meilin Jiang (Duke)
- Siyeon Kim (UNC Chapel Hill)
- Duyeol Lee (UNC Chapel Hill)
- Gang Li (UNC Chapel Hill)
- Rui Li (NCSU)
- Na Lin (UNC Chapel Hill)
- Jitong Lou (UNC Chapel Hill)
- Vikram Patil (NCSU)
- Sumit Kumar Sar (UNC Chapel Hill)
- Junhui Wang (IBM)
- Wenbo Wang (UNC Chapel Hill)
- Zhe Wang (NCSU)
- Adane Fekadu Wogu (UNC Chapel Hill)
- Hang Yu (UNC Chapel Hill)
- Beibo Zhao (UNC Chapel Hill)
- Xianglin Zhao (Duke)
Transportation and Parking

Local Time

Raleigh is in Eastern Time Zone.

Location

The 2019 Applied Statistics Symposium by the International Chinese Statistical Association (ICSA) will be held at Raleigh Convention Center (https://www.raleighconvention.com). The address is 500 S Salisbury St, Raleigh, North Carolina, USA, 27601.

Airports

The closest airport is Raleigh-Durham International Airport (https://www.rdu.com).

Public Transportation to Raleigh Convention Center

- Public transportation from Raleigh-Durham airport to the hotel:
  
  GoTriangle's Route 100 bus picks up and drops off at the airport from Downtown Raleigh's GoRaleigh Station every 30 minutes through 6:30 PM and then hourly through the evening. The GoRaleigh Station is a convenient 3 1/2 blocks from the RCC.

Parking and Driving

- View a map of parking decks in Downtown Raleigh:
  
  https://www.downtownraleigh.org/getting-around/parking

- The closest public parking decks to the Raleigh Convention Center are located at
  
  - Lenoir Street between Salisbury Street and Fayetteville Street
  - Lenoir Street between Salisbury and McDowell Street
  - South Street between Salisbury and McDowell Street
  - Davie Street between McDowell and Dawson Street
  - Cabarrus Street between McDowell and Dawson Street
  - Salisbury Street between Cabarrus and Davie Street

- For more information on driving to the center, please refer to the RCC website:
  
  https://www.raleighconvention.com/directions
Downtown Dining Information

View a map of dining information in Downtown Raleigh:

https://www.visitraleigh.com/DTdining

Wi-Fi Information

RCC provides the complimentary Wi-Fi, which is suitable for light surfing of the web and for checking email. Attendees may access the complimentary Wi-Fi by using the Raleigh Convention Center network or Legacy Devices network. Once you select a network you will accept the terms and conditions and you will be logged into the network. No password is needed.
Opening Mixer:

Location and Time: Raleigh Convention Center Room 402, June 9 (Sunday), 6:30 pm – 9:00 pm

Banquet:

Location and Time: Raleigh Convention Center Ballroom C, June 11 (Tuesday), 6:45 pm – 9:00 pm
Banquet speaker: David Banks, PhD,
    University Professor and Director of the Statistical and Applied Mathematical Sciences Institute,
    Department of Statistical Science at Duke University
### Sunday June 9, 2019

<table>
<thead>
<tr>
<th>Time</th>
<th>Room</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:30 am-8:30 am</td>
<td>RCC Main Lobby</td>
<td>Breakfast</td>
</tr>
<tr>
<td>8:30 am-12:30 pm</td>
<td>See program</td>
<td>Morning short course</td>
</tr>
<tr>
<td>1:30 pm-5:30 pm</td>
<td>See program</td>
<td>Afternoon short course</td>
</tr>
<tr>
<td>6:30 pm-9:00 pm</td>
<td>RCC Room 402</td>
<td>Mixer</td>
</tr>
</tbody>
</table>

### Monday June 10, 2019

<table>
<thead>
<tr>
<th>Time</th>
<th>Room</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 am - 9:10 pm</td>
<td>RCC Ballroom C</td>
<td>Open remarks</td>
</tr>
<tr>
<td>9:10 am - 10:10 am</td>
<td>RCC Ballroom C</td>
<td>Keynote I</td>
</tr>
<tr>
<td>10:10 am - 10:30 am</td>
<td>See program</td>
<td>break</td>
</tr>
<tr>
<td>10:30 am - 12:10 pm</td>
<td>See program</td>
<td>Parallel Sessions</td>
</tr>
<tr>
<td>2:00 pm - 3:40 pm</td>
<td>See program</td>
<td>Parallel Sessions</td>
</tr>
<tr>
<td>3:40 pm - 4:00 pm</td>
<td>See program</td>
<td>break</td>
</tr>
<tr>
<td>4:00 pm - 5:40 pm</td>
<td>See program</td>
<td>Parallel Sessions</td>
</tr>
</tbody>
</table>

### Tuesday June 11, 2019

<table>
<thead>
<tr>
<th>Time</th>
<th>Room</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 am - 9:00 am</td>
<td>RCC Ballroom C</td>
<td>Keynote II</td>
</tr>
<tr>
<td>9:10 am - 10:10 am</td>
<td>RCC Ballroom C</td>
<td>Keynote III</td>
</tr>
<tr>
<td>10:10 am - 10:30 am</td>
<td>See program</td>
<td>break</td>
</tr>
<tr>
<td>10:30 am - 12:10 pm</td>
<td>See program</td>
<td>Parallel Sessions</td>
</tr>
<tr>
<td>2:00 pm - 3:40 pm</td>
<td>See program</td>
<td>Parallel Sessions</td>
</tr>
<tr>
<td>3:40 pm - 4:00 pm</td>
<td>See program</td>
<td>break</td>
</tr>
<tr>
<td>4:00 pm - 5:40 pm</td>
<td>RCC Ballroom C</td>
<td>Parallel Sessions</td>
</tr>
<tr>
<td>6:45 pm - 9:00 pm</td>
<td>RCC Ballroom C</td>
<td>Banquet</td>
</tr>
</tbody>
</table>

### Wednesday June 12, 2019

<table>
<thead>
<tr>
<th>Time</th>
<th>Room</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30 am - 10:10 am</td>
<td>See program</td>
<td>Parallel Sessions</td>
</tr>
<tr>
<td>10:10 am - 10:30 am</td>
<td>See program</td>
<td>break</td>
</tr>
<tr>
<td>10:30 am - 12:10 pm</td>
<td>See program</td>
<td>Parallel Sessions</td>
</tr>
</tbody>
</table>
Lilly Yue, Ph.D.

Lilly Yue, Ph.D., is Deputy Director, Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration (FDA), and oversees pre-market statistical reviews of medical devices and diagnostics. She serves in the prestigious Senior Biomedical Research Service in the U.S. Department of Health and Human Services. Her research interests include the observational study design and data analysis using propensity score methodology in the regulatory settings, subgroup analysis, and missing data handling. She has served as Associate Editor for the *Journal of Biopharmaceutical Statistics* and *Pharmaceutical Statistics*. She was the 2012 President of FDA Statistical Association and Co-Chair of 2013 FDA-Industry Statistics Workshop. She also served on the Board of Directors of International Chinese Statistical Association (2015 - 2017). She earned a B.S. degree in Mathematics, two M.S. degrees in Stochastic Operations Research and Mathematical Statistics, respectively, and a Ph.D. in Mathematical Statistics from Texas A&M University. She is a Fellow of the American Statistical Association and recipient of the FDA Center/ORA Scientific Achievement Award - Excellence in Analytical Science.

Title: Leveraging Real-World Data for Regulatory Decision-Making: Propensity Score-Based Patient Selection and Augmentation in Investigational Clinical Studies

Location and Time: RCC Ballroom C, June 10 (Monday), 9:10 am – 10:10 am

Keynote Host: Lanju Zhang, Ph.D., AbbVie Inc.

Abstract:
In medical product development, there has been an increased interest in utilizing real-world data which have become abundant owing to advances in biomedical science, information technology and engineering. High-quality real-world data may be utilized to generate real-world evidence for regulatory or healthcare decision-making. This presentation will discuss two propensity score-based methods for leveraging patients from a real-world data source: one for constructing a control group for a comparative clinical study, and the other for augmenting a single-arm clinical study. The proposed approaches use the propensity score methodology to identify real-world patients who are similar to those enrolled into the investigational clinical study in terms of baseline characteristics. Either frequentist or Bayesian method can then be applied to outcome data analysis, with the option of down-weighting information from real-world data source. Examples based on pre-market regulatory review experience are provided to illustrate the implementation of the proposed approaches.
Keynote Speaker

Stephen J. Ruberg, Ph.D.

Stephen J. Ruberg, Ph.D., is President, Analytix Thinking, LLC. Steve was in the pharma industry for 38 years where he worked in all phases of drug development and commercialization – from R&D to Business Analytics. He retired from Lilly at the end of 2017 where he spent the last 10 years of his career there as a Distinguished Research Fellow and the Scientific Leader for Advanced Analytics. Since then, he has formed his own consulting company - Analytix Thinking. He is a Fellow of the ASA since 1994, and an elected Fellow of ISI. He has published widely and lectured internationally on a variety of topics relevant to pharmaceutical research. In particular, he has helped pioneer many areas of data and statistics including:

- Publications on dose response design/analysis, starting with a JASA paper in 1989;
- Analysis of drug stability studies, which is the basis of ICH-Q1E "Evaluation for Stability Data;"
- Co-authorship of ICH-E9, Statistical Principles for Clinical Trials;
- Collaboration on understanding QTc intervals and designed the first thorough QTc study;
- Establishment CDISC and was its first Chairman of the Board for 6 years;
- Creating strategies for electronic medical records (EMRs) where he served on an Advisory Board to the Secretary of Health and Human Services in the George W. Bush Administration;
- Research on personalized medicine and biomarker research through subgroup identification, Bayesian methods and statistical analysis of SNPs;
- And most recently estimands and causal inference.

Title: Are you sure? The power of statistical thinking

Location and Time: RCC Ballroom C, June 11 (Tuesday), 8:00 am – 9:00 am

Keynote Host: Haoda Fu, Ph.D., Eli Lilly and Company

Abstract: The world is exploding with data, and the application of ‘analytics’ is growing at a commensurate rate. The question remains: “Are we using the right data, and are we applying smart analytics?” This talk will explore three topics that are major trends in our current scientific environment – big data, inference and validated analysis. Pros and cons of each of these will be explored, and fundamental concepts elucidated to make our statistical thinking clearer and more precise. Statisticians must grasp these trends, understand the fundamental issues underlying them and do a better job communicating how to do smart analytics. This can influence policy decisions by governments and have major societal impact, all for the betterment of mankind. This will also require statisticians to reinvent themselves to play a much broader role in the most difficult problems of our time.
Marie Davidian, Ph.D.

Marie Davidian is J. Stuart Hunter Distinguished Professor of Statistics at North Carolina State University (NCSU), and Adjunct Professor of Biostatistics and Bioinformatics at Duke University. She is a Fellow of the American Statistical Association, the Institute of Mathematical Statistics, and the American Association for the Advancement of Science, and is an elected member of the International Statistical Institute. Marie has served as chair of grant review panels for the National Institutes of Health, as Coordinating and Executive Editor of Biometrics, as a member of US Food and Drug Administration Advisory Committees, and as president of the American Statistical Association and of the Eastern North American Region of the International Biometric Society. Marie’s research interests include design and analysis clinical trials and observational studies, statistical inference in the presence of missing or mismeasured data, analysis of longitudinal data, and causal inference and dynamic treatment regimes. She is Principal Investigator for a P01 Program Project grant, “Statistical Method for Cancer Clinical Trials,” awarded by the National Cancer Institute to a consortium of NCSU, Duke, and the University of North Carolina at Chapel Hill; and she has been Director of the NCSU-Duke Summer Institute in Biostatistics (SIBS) program, sponsored by the National Heart, Lung, and Blood Institute to encourage US undergraduates to pursue graduate training in biostatistics since 2004. She received the 2007 Janet L. Norwood Award for Outstanding Achievement by a Woman in the Statistical Sciences; the 2009 George W. Snedecor and the 2011 F.N. David Awards presented by the Committee of Presidents of Statistical Societies; the 2018 ASA Founders Award; and the 2010 NCSU Alexander Quarles Holladay Medal for Excellence, the highest honor the University bestows on a faculty member.

Title: Modernizing Statistics Through Treatment Regimes: A Review

Location and Time: RCC Ballroom C, June 11 (Tuesday), 9:10 am – 10:10 am

Keynote Host: Wenbin Lu, Ph.D., North Carolina State University

Abstract:
Statistics has played and will continue to play a fundamental role in health sciences research and in particular in the design and analysis of studies of treatments for chronic diseases and disorders, and the current focus on precision medicine presents numerous opportunities for statistics. The past decade has seen an explosion of statistical methodological research for discovery of optimal treatment regimes from data; a treatment regime is a set of sequential decision rules that maps individual patient characteristics to recommended treatments from among the available, feasible options. Precision medicine seeks to make clinical decision-making evidence-based; thus, methodology for estimation of treatment regimes from data has a significant role to play in advancing this objective. This talk will review the evolution of methodology for treatment regimes and advocate for broader adoption of treatment regimes in practice.
David Banks is currently the Director of the Statistical and Applied Mathematical Sciences Institute, and a professor in the Dept. of Statistical Science at Duke University. He has held previous positions at UC Berkeley, the University of Cambridge, Carnegie Mellon, the National Institute of Standards and Technology, the US Dept. of Transportation, and the FDA. He obtained his PhD in 1984 at Virginia Tech, and has served as editor of JASA and Statistics and Public Policy. He is interested in dynamic text networks, risk analysis, agent-based models, biosurveillance, computational advertising, and the estimation of cryptic populations.

Location and Time of the Speech: RCC Ballroom C, June 11 (Tuesday), 7:30 pm
Student Paper Awards

Jiann-Ping Hsu Pharmaceutical and Regulatory Sciences Student Paper Award

Xinzhou Guo, University of Michigan
— Title: A Quantitative Assessment of Risk for Subgroup Pursuit in Clinical Trials
— Time: June 11 (Tuesday) 4:00 pm – 5:40 pm
— Session 76: The Jiann-Ping Hsu Invited Session on Biostatistical and Regulatory Sciences

ICSA Student Paper Awards

Zhou Lan, North Carolina State University
— Title: Geostatistical Modeling of Positive Definite Matrices Using the Spatial Wishart Process
— Time: June 10 (Monday) 4:00 pm – 5:40 pm
— Session 39: Student Paper Award Session

Xiangyu Liu, University of Texas Health Science Center at Houston
— Title: Semiparametric Modelling and Estimation of the Global Percentile Outcome
— Time: June 10 (Monday) 4:00 pm – 5:40 pm
— Session 39: Student Paper Award Session

Jitong Lou, University of North Carolina at Chapel Hill
— Title: Integrative Analysis of Irregularly Measured and Mixed-Type Biomarkers in Electronic Health Records Method with Statistical Guarantees
— Time: June 10 (Monday) 4:00 pm – 5:40 pm
— Session 39: Student Paper Award Session

Jennifer Starling, University of Texas at Austin
— Title: BART with Targeted Smoothing: An Analysis of Patient-Specific Stillbirth Risk
— Time: June 10 (Monday) 4:00 pm – 5:40 pm
— Session 39: Student Paper Award Session

Tao Sun, University of Pittsburgh
— Title: GWAS-based AMD Progression Using a Copula Semiparametric Model
— Time: June 10 (Monday) 4:00 pm – 5:40 pm
— Session 39: Student Paper Award Session
SC01: A Variety of Mixed Models

Location and Time: RCC 305A, June 9 (Sunday), 8:30 am – 12:30 pm

David A. Dickey, North Carolina State University

Abstract:
Mixed models are those with fixed and random effects. In ordinary mixed models, one estimates the fixed effects using estimated generalized least squares where the variance-covariance matrix of the data is estimated as part of maximum likelihood or REML (Restricted, or Residual, Maximum Likelihood) algorithm. After reviewing how to distinguish random from fixed effects, this course will describe the overall methodology and show several examples of its application including random coefficient models, repeated measures and hierarchical models. A review of nonlinear models is included and the additional complexities arising from the inclusion of random effects illustrated. A third type of model, the generalized linear mixed model, is discussed with examples. Such a model arises when the response is not normally distributed but rather is in the exponential family of distributions. Outstanding examples of the exponential family are the binomial and Poisson distributions. Emphasis is on concepts, examples, when to apply each type of model, and how to interpret each. Examples use SAS™ but the ideas presented are independent of software.

TM SAS is the registered trademark of SAS Institute, Cary, NC

About the instructor:
Dr. David A. Dickey is W. N. Reynolds Professor (emeritus) of Statistics at North Carolina State University. He is known for the Dickey-Fuller test for unit roots in time series. He is a Fellow of the American Statistical Association. He has spoken at the ASA’s JSM, ASQ, and CSP meetings and many times at SAS Global Forum and regional SAS Users’ Group meetings. Dickey has co-authored several books and dozens of papers. He was major advisor to 16 PhD students at NCSU and served on hundreds of graduate student committees across campus. Dickey is a member of NCSU’s Academy of Outstanding Teachers and Academy of Outstanding Faculty Engaged in Extension. He received the D.D. Mason Faculty Award in 1986 and 2018 and the Outstanding Extension Service Award in 2007. Dickey was a founding faculty member of NCSU’s Institute for Advanced Analytics, holds an associate appointment in Economics, and is a member of the Financial Math faculty. He taught at Randolph Macon College and The College of William and Mary in Virginia for 3 years before earning his PhD in 1976 under Wayne Fuller and spending the next 43 year at NC State.

SC02: Longitudinal Data Analysis and Latent Growth Curve Modelling in Public Health

Location and Time: RCC 305A, June 9 (Sunday), 1:30 pm – 5:30 pm

Din Chen, University of North Carolina at Chapel Hill

Abstract:
Longitudinal data are very commonly collected by several time points which produced longitudinal data to evaluate the growth curve. However, the longitudinal data are temporally correlated which directly violate the fundamental assumption of independence in typically regression modelling and therefore erroneous conclusions and recommendations could be made if the data are not analyzed appropriately. Therefore, an appropriate analysis of longitudinal data is important to capture the intra-individual growth changes and inter-individual variabilities. This workshop is then designed to show how to do longitudinal data analysis using R package “nlme/lmerTest” and latent growth curve modelling using Mplus. Real data on study of 405 Hong Kong Chinese women...
who underwent cancer surgery will be used as a real example in the class to model the evidence of rate change in their mood and social adjustment at 1, 4, and 8 months post-surgery.

About the instructor:
Dr. Din Chen is a Fellow of ASA. He is now the Wallace H. Kuralt distinguished professor in biostatistics, University of North Carolina at Chapel Hill. He was a professor in biostatistics at the University of Rochester and the Karl E. Peace endowed eminent scholar chair in biostatistics at Georgia Southern University. Professor Chen is also a senior statistics consultant for biopharmaceuticals and government agencies with extensive expertise in clinical trials and bioinformatics. He has more than 150 refereed professional publications and co-authored/co-edited 23 books on randomized clinical trials, statistical meta-analysis, public health statistical methods, causal inferences and statistical Monte-Carlo simulation and public health applications. Professor Chen has more than 30 years teaching and research experience in statistics/biostatistics. He has been invited to give short courses at JSM, CSP, ICSA, Biopharmaceutical Applied Statistics Symposium, Deming Conferences. He was honored with the "Award of Recognition" by the Deming Conference Committee for highly successful biostatistics workshop tutorials. This course on latent growth curve modelling will be of interest to even broad audiences.

SC05: Re-Sampling: Theory, Uses, Novel Uses, and Mis-uses

Location and Time: RCC 306A, June 9 (Sunday), 8:30 am – 12:30 pm

Richard Simon, formerly National Cancer Institute and currently R Simon Consulting

Abstract:
The fields of biology and medicine have seen the introduction of modern technologies for measuring biological processes within cells. Instead of having to develop tailored assays for each gene or protein of interest, the new assays provide genome-wide information about gene expression, epigenetic silencing as well as point mutations and copy number genomic variations. These development in technology have led to new biomedical problems and the development of new statistical methods for the analysis of p>n data where the number of variables is greater than the number of cases. Re-sampling methods have seen increased use as replacements for model-based methods which may give misleading results in such circumstances. It has also resulted in new emphases on classification and prediction and re-sampling methods have found new uses for the evaluation of prognostic and predictive classifiers. In this half day course, we will explore a variety of uses of re-sampling including some novel uses for model evaluation. We will start with a review of some theory for re-sampling methods and addressing some views on the validity of such methods. This half day course is intended for biostatisticians as most examples and novel uses will be from that area. No specific pre-requisite is required except some experience with R. The teaching plan will include lectures, and the illustration of worked examples using software distributed to the participants.

About the instructor:
Dr. Richard Simon retired recently from his four-years’ service at the National Cancer Institute, serving most recently as Director of the Biometric Research Program and Chief of the Computational and Systems Biology Branch. He now has his own consulting practice: http://rsimon.us. He is the author of the book: “Genomic Clinical Trials and Predictive Medicine.” He also coauthored two other books: “Design and Analysis of DNA Microarray Investigations” and “Design and Analysis of Clinical Trials for Predictive Medicine”. He is a former member of the FDA Oncologic Drug Advisory Committee and a Fellow of the American Statistical Association. He is the recipient of the 2013 Karl Peace award “for contributions that have played a pivotal role in bridging the gap among statistics, clinical research, and translational medicine to improve human health” and the recipient of the 2017 Marvin Zelen award for leadership in Statistical Science from the Department of Biostatistics at Harvard University.
SC06: Causal Analysis of Observational Data Using SAS/STAT® Software

Location and Time: RCC 306A, June 9 (Sunday), 1:30 pm – 5:30 pm

Michael Lamm and Clay Thompson, SAS Institute Inc.

Abstract:
The analysis of data from observational studies is an increasingly common task faced by data scientists and applied statisticians. Like analyzing data from randomized experiments, drawing causal inferences from observational studies requires careful consideration of the underlying data generating process and research design. Indeed, because of the presence of confounding variables, these requirements are even more critical for observational studies. Hence, specialized analytic tools and estimation methods have been developed for estimating causal effects from observational data. This presentation demonstrates the uses of these tools and methods. This tutorial begins by reviewing the theory of causal effect estimation and discussing the difficulties that arise in the absence of randomization. It then introduces propensity score matching, inverse probability weighting, and doubly robust methods for estimating a treatment effect from observational data and explores the use of directed causal graphs to represent the data generating process. Causal graph analysis identifies sources of association and bias and evaluates whether a study design supports the estimation of a causal effect with a valid causal interpretation. Several examples, worked out in detail, illustrate approaches for constructing and evaluating the underlying models, comparing the estimation methods, and examining the assumptions required for the identification and estimation of treatment effects. Each example emphasizes elements of a rigorous and comprehensive workflow for causal analysis from observational or imperfectly randomized studies. The examples in this tutorial use SAS/STAT® software. No prior experience with these estimation and graphical methods is assumed.

About the instructors:
Dr. Michael Lamm is a Senior Research Statistician Developer in the Advanced Regression Research Department at SAS Institute. Among his responsibilities are the development of software for longitudinal data analysis and causal inference. Before coming to SAS, he received his PhD in statistics and operations research from the University of North Carolina at Chapel Hill. Dr. Clay Thompson is a Senior Research Statistician Developer in the Multivariate Models Research Department at SAS Institute, where he develops algorithms and software for the analysis of causal effects using graphical models. Before joining SAS, he worked as a quantitative systems pharmacologist in the pharmaceutical industry. He received a PhD in applied mathematics from North Carolina State University.

SC07: Advanced Topics in Propensity Score Methods

Location and Time: RCC 306B, June 9 (Sunday), 8:30 am – 12:30 pm

Wei Pan, Duke University

Abstract:
Propensity score methods have become a widespread practice in observational studies to reduce selection bias. However, researchers still often encounter in-depth methodological challenges in practice. For example, how to check assumptions? How to select covariates? How to deal with multiple treatments? How to deal with complex data such as clustered, longitudinal, and survey data? How to conduct outcome analysis using propensity score methods? How to conduct sensitivity analysis after applying propensity score methods? This course will build upon the fundamental principle and concept of propensity score methods to tackle these in-depth methodological challenges along with relevant R packages, such as twang, Zelig, and rbounds. Through lectures on the advanced topics in propensity score methods and hands-on activities for using the R packages, this
course will help researchers better understand propensity score methods and, therefore, improve the validity of their observational studies. Instructions for downloading and installing the R packages with examples of real-world data will be provided to participants in advance through a course website. A working knowledge of propensity score methods is required. Participants are encouraged to bring their own computers for hands-on activities on the data provided in the course, and they are also welcomed to work on their own real-world data.

**About the instructor:**
Dr. Wei Pan is an Associate Professor of Research Design and Statistics at the Duke University School of Nursing. His research work focuses on causal inference (propensity score methods, resampling, power and effect size), advanced statistical modeling (multilevel, structural, and longitudinal), meta-analysis, psychometrics, and their applications in the social, behavioral, and health sciences. He has published numerous referred journal articles on both methodological and applied research studies, including an edited book on propensity score methods entitled, “Propensity Score Analysis: Fundamentals and Developments.” He has been invited or selected to offer more than ten professional development courses and workshops on propensity score methods at annual conferences of various professional organizations, such as the American Statistical Association (JSM and CSP), the American Public Health Association, the International Chinese Statistical Association (ICSA), the American Educational Research Association, the American Evaluation Association, and the South African Statistical Association.

**SC08: Power Prior: Incorporating Historical Data for Bayesian Inference and Designs of Clinical Trial**

Location and Time: RCC 306B, June 9 (Sunday), 1:30 pm – 5:30 pm

Ming-Hui Chen, University of Connecticut

**Abstract:**
The power prior has been widely used in many applications covering a variety of disciplines. The power prior is intended to be an informative prior constructed from historical data. This short course provides a comprehensive overview of the power prior and its applications to date. The course starts with a brief introduction of the basic formulation of power prior and then presents various variations of power prior, including the full power prior, the normalized power prior, the partial discounting power prior, and the partial borrowing power prior. The theory of the power prior and its properties, including the theoretical justification, connections to hierarchical models, and frequentist properties of the posterior estimates using power priors in linear models as well as generalized linear models will be reviewed. The determination of a guide value of the power parameter will be discussed. A revised information gain measure will be introduced and illustrated on how to quantify the information gain or loss while incorporating historical data. A detailed analysis of benchmark dose toxicology data will be presented.
The course also covers an important application in the design of clinical trials from the Bayesian perspective. A general Bayesian methodology for the design of non-inferiority clinical trials with a focus on controlling type I error and power will be presented. Bayesian methods are then applied to the design of a non-inferiority medical device clinical trial with historical data from previous trials to demonstrate superiority of the Bayesian methods in sample size reduction.

**About the instructor:**
Dr. Ming-Hui Chen is currently Professor and Head of the Department of Statistics at the University of Connecticut (UConn). He was elected to Fellow of the International Society for Bayesian Analysis in 2016, Fellow of the Institute of Mathematical Statistics in 2007, and Fellow of the American Statistical Association in 2005. He has published over 375 statistics and biostatistics methodological and medical research papers in mainstream statistics, biostatistics, and medical journals. He has also published five books, including two advanced graduate-level books on Bayesian survival analysis and Monte Carlo
methods in Bayesian computation. He served as President of the International Chinese Statistical Association (ICSA) in 2013, Program Chair and Publication Officer of SBSS of the American Statistical Association (ASA) and the ASA Committee on Nomination for 2016-2017 to nominate candidates for ASA President/Vice President. Currently, he serves as Co Editor-in-Chief of Statistics and Its Interface and an Associate Editor of JASA, JCGS, and LIDA.

SC10: Statistical Analysis of Network Data

Location and Time: RCC 304, June 9 (Sunday), 8:30 am – 12:30 pm and 1:30 pm – 5:30 pm

Eric Kolaczyk, Boston University

Abstract:
Networks have permeated everyday life through everyday realities like the Internet, social networks, and viral marketing. Their use has become especially prevalent in the biological and life sciences, particularly in computational biology and neuroscience. Accordingly, network analysis is an important growth area in the quantitative sciences, with roots in social network analysis going back to the 1930s and graph theory going back centuries. Measurement and analysis are integral components of network research, and statistical methods therefore play a critical role in network analysis. This course will provide a broad treatment of foundational topics relevant to statistical analysis of network data across the disciplines. Material will be organized according to a statistical taxonomy, with presentation entailing a conscious balance of conceptual and technical aspects. Additionally, practical application of network analysis will be demonstrated in the context of the R software environment. The focus in the morning will be on manipulation, visualization, and descriptive analysis of network data. In the afternoon, we will draw on topics from network sampling and inference, the modeling of networks and network-indexed processes, and networked experiments. Specific examples of network analysis will be taken from a variety of domain areas, with emphasis on computational biology and neuroscience and on social networks.

About the instructor:
Dr. Eric Kolaczyk is a tenured professor and director of the program in statistics in the Department of Mathematics and Statistics at Boston University, where he is also a Data Science Faculty Fellow and an affiliated faculty member in the program in bioinformatics, the program in computational neuroscience, and the division of systems engineering. His major research interests for 15 years have revolved around the statistical analysis of network-indexed data and include both the development of basic methodology and inter-disciplinary work with collaborators in bioinformatics, computer science, geography, neuroscience, and sociology. He has authored a book: “Statistical Analysis of Network Data: Methods and Models” and coauthored another book: “Statistical Analysis of Network Data with R”. Dr. Kolaczyk has served as associate editor on several journals, including JASA and JRSS-B. He has also served as (co)organizer for workshops focused on networks and network data, including as lead organizer for a year-long program at SAMSI in 2010-11. He is an elected fellow of the American Statistical Association (ASA) and the Institute for Mathematical Statistics (IMS), as well as of the American Association for the Advancement of Science (AAAS). He is also an elected senior member of the Institute for Electrical and Electronics Engineers (IEEE), and an elected member of the International Statistical Institute (ISI).

SC11: Data Science, Big Data, and Deep Learning for Statistician

Location and Time: RCC 402, June 9 (Sunday), 8:30 am – 12:30 pm and 1:30 pm – 5:30 pm

Hui Lin, Netlify
Ming Li, Amazon.com Inc.

Abstract:
With recent big data, data science and deep learning
Short Courses

revolution, enterprises ranging from FORTUNE 100 to startups across the world are hungry for data scientists and machine learning scientists to bring actionable insight from the vast amount of data collected. In the past a couple of years, deep learning has gained traction in many application areas and it becomes an essential tool in data scientist’s toolbox. In this course, participant will develop a clear understanding of the big data cloud platform, technical skills in data sciences and machine learning, and especially the motivation and use cases of deep learning through hands-on exercises. We will also cover the “art” part of data science and machine learning to guide participants to learn typical agile data science project flow, general pitfalls in data science and machine learning, and soft skills to effectively communicate with business stakeholders. The big data platform, data science, and deep learning overviews are specifically designed for audience with statistics education background. This course will prepare statisticians to be successful data scientists and machine learning scientist in various industries and business sectors with deep learning as focuses.

About the instructor:
Dr. Hui Lin is leading and building data science department at Netlify since 2018. Before Netlify, she was a Data Scientist at DowDuPont. She was a leader in the company of applying advanced data science to enhance Marketing and Sales Effectiveness. She provided data science leadership for a broad range of predictive analytics and market research analysis from 2013 to 2018. 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.

Dr. Ming Li is currently a Research Scientist at Amazon. He organized and presented 2018 JSM Introductory Overview Lecture: Leading Data Science: Talent, Strategy, and Impact. He was 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.

SC12: Advancing Clinical Development through Precision Medicine and Innovative Designs: Concepts, Rationale, and Case Studies

Location and Time: RCC 306C, June 9 (Sunday), 8:30 am – 12:30 pm and 1:30 pm – 5:30 pm

Sandeep M Menon, Pfizer Inc. / Boston University/Tufts University School of Medicine
Weidong Zhang, Pfizer Inc.

Abstract:
Precision medicine has paved the way for a new era of delivering tailored treatment options to patients according to their biological profiles. Advancement of the biotechnologies such as next generation sequencing technology (NGS) and other omics technologies have enabled us to interrogate a patient’s many molecular biomarkers and associate them with disease and drug responses. In addition, incorporation of biomarker information in the innovative clinical trial design has presented drug developers unprecedented opportunities to bring a successful drug to patients in needs. The first part of this course will focus on the concept of precision medicine, biomarker discovery and its application in clinical development. Comprehensive review of omics data and major technologies will be presented. Statistical considerations and challenges such as data normalization, dimension reduction and biomarker threshold development will be discussed. Strategies in
Bayesian framework leveraging historical biomarker data for quantitative decision making in early clinical trials will also be presented in detail. The second part of this course will focus on the strategy of the study design that is important to critically determine biomarker performance, reliability and eventually regulatory acceptance. A general overview of the concept and statistical methodologies and designs related to precision medicine will be presented. Specifically, we will discuss a variety of designs including adaptive designs available at our disposal and its merits and limitations.

About the instructor:
Dr. Sandeep Menon is currently the Vice President and the Head of Early Clinical Development Statistics at Pfizer Inc. and holds adjunct faculty positions at Boston University and Tufts University School of Medicine. He is an elected fellow of American Statistical Association. He leads an organization of early development statisticians, clinicians and pharmacologists globally and is the executive member of the early clinical development leadership team and global clinical leadership team. He is internationally known for his technical expertise especially in adaptive designs, precision medicine, multi-regional trials, and small populations. He has participated in the core review of draft version of the regulatory guidance documents. He has co-authored and co-edited numerous books and contributed to influential papers in this area. He has taught short courses internationally and is a regular invited speaker in academia, FDA, Industry forums and Business Management institutes. He is the co-author and co-editor of the books titled “Clinical and Statistical Considerations in Personalized Medicine”, “Modern Approaches to Clinical Trials Using SAS: Classical, Adaptive, and Bayesian Methods” and “Biosimilars – Clinical Development.” Dr. Menon completed his master’s and Ph.D. in Biostatistics at Boston University and research assistantship at Harvard Clinical Research Institute. He has received several awards for academic excellence.

Dr. Weidong Zhang is a Senior Director in the Statistical Research and Innovation department at Pfizer. His responsibilities include statistical methodology development for precision medicine and providing scientific leadership and consultation in clinical biomarker strategy to senior Pfizer management and portfolios spanning from target discovery through proof-of-concept, and late phase studies across multiple therapeutic areas including oncology, immunology and inflammation, rare disease, and cardiovascular and metabolism. His research interest focuses on developing new statistical methods in biomarker discovery and precision medicine studies using high throughput omics data generated from cutting edge technologies including next-generation sequencing technology. He has taught short courses for ASA Biopharmaceutical Section Regulatory-Industry Statistics Workshop, the Biopharmaceutical Applied Statistics Symposium (BASS), and the International Chinese Statistics Association (ICSA) Applied Statistics Symposium. He obtained his PhD degree in Statistical Genetics and MS degree in Statistics from the University of Wisconsin-Madison.
2019 ICSA Applied Statistics Symposium Sponsors

The 2019 ICSA Applied Statistics Symposium Program Committees gratefully acknowledge the generous support of our sponsors below.

**Gold Sponsors**

- FMD K&L
- TechData
- Boehringer Ingelheim

**Silver Sponsors**

- SAS
- SANOFI PASTEUR
- Pfizer
- MERCK
- abbvie

The ICSA 2019 Symposium sponsors currently have the following statistics related openings.
Sponsors and Career Services

Biostatistics
- Randomization
- IDMC/DSMB
- SAP
- Protocol/CSR
- ISS/ISE

Data Management
- EDC Database Development
- Data Review and Query Management
- Dictionary Coding

Statistical Programming
- SDTM/ADaM
- Table/Figure/Listing
- Legacy Data Conversion
- SDRG/ADRG
- eSubmission Ready

Drug Safety
- Case Entry
- Narrative Writing
- SAE Reconciliation
- Global Case Submission

Proud To Be ICSA Sponsor Again!
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.
The qualified candidate will be hired at the appropriate level commensurate with education/experience.

Description:
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.

As an employee of Boehringer Ingelheim, you will actively contribute to the discovery, development and delivery of our products to our patients and customers. Our global presence provides opportunity for all employees to collaborate internationally, offering visibility and opportunity to directly contribute to the companies' success. We realize that our strength and competitive advantage lie with our people. We support our employees in a number of ways to foster a healthy working environment, meaningful work, diversity and inclusion, mobility, networking and work-life balance. Our competitive compensation and benefit programs reflect Boehringer Ingelheim’s high regard for our employees.

Duties & Responsibilities:

- Perform duties of a Trial Statistician to support complex clinical trials within national or international development projects or for marketed products as required. Collaborate with members of Clinical Research and Marketing, Trial Clinical Monitor and trial teams incl. pharmacokineticist in planning clinical trials and protocols conforming to company and regulatory agency guidelines and/or marketing and publication strategies.
- Act as Project Statistician for early projects, backup projects, or projects with established BI experience.
- Support Project Statisticians of high profile international projects in their responsibilities, especially in their statistical responsibilities in the planning and preparation of regulatory submissions and contribute to efforts on cross-trial planning and harmonization.
- Analyze data from phase I to IV trials incl. responsibility for program validation. Perform exploratory analyses in collaboration with the Project Statistician.
- Prepare accurate, high quality reports of complex clinical trials for registration of drugs and biologics, publications and management.
- Prepare specifications for data analyses by outside vendors as required. Assure compliance with the specifications by reviewing the vendors’ products.
- Participate on assigned international teams to promote harmonization efforts for clinical drug development.
- Support management in resource planning and tracking for assigned trials and projects.
- Act as a team leader for a complex project or mega-trial who ensures team members adhere to the SOPs, guidelines and local working instructions.
- Attends all the meetings related to the trial/project needing a statistical input (or send delegates) and send minutes to the team members.
- Assist the Head of programming and Head of statistics with the working of vendors, contractors in establishing procedures for programming and validating statistical analysis (writes the scope of work, prepare documents to be sent to the contract research organization (CRO). From a statistical perspective, is the primary contact for CROs (programming validation)
Acts as a Trial or Project Statistician (TSTAT or PSTAT) for the trial/project. In particular, develops report and programming specifications in a Trial Statistical Analysis Plan (TSAP) and update this document as often as needed.

Collaborate with the programming group and the data management group to submit very detailed timelines for the trial/project to clinicians. Meeting agreed upon timelines are essential to the success of the clinical trial/project team objectives.

From a statistical perspective, is the primary contact for medical writing (or choose a designee).

Ensures that protocol objectives are met and project standards are maintained (also responsible to update the project statistical analysis plan when necessary).

Ensures achievement of major statistical deliverables and milestones in coordination with other functions including Clinical Research, Safety, Statistical Programming, Data Management and Medical Writing.

Provides and organizes statistical support for regulatory meetings, questions and submissions.

Informs management that hours allocated to TSTAT by the capacity algorithm should be updated (follow up the amount of hours entered in the time recording system and related to the trial/project).

Ensures efficient work within the team by setting priorities and avoiding overlaps between team members.

Works directly and proactively with the Trial, Project or Substance team.

Assumes responsibility for the coordination of all relevant statistical activities for a Trial, Project or Substance.

Receives broad operating instructions in performing a majority of duties from manager and keeps manager abreast of programming status. Alerts management in case of resources issues, timelines problems or conflicts within the team.

In collaboration with the statistical expert group member, maintain expertise in therapeutic area, by keeping abreast of new publications with the purpose of increasing the overall efficiency or effectiveness in the department.

Sr. Biostatistician Requirements:

- M.S. in statistics, biostatistics, or biometry with three years of experience in: designing, conducting, analyzing and/or presenting routine trials/studies, and working with a team to apply statistical methodology to a research question.
- Or a Ph.D. in statistics, biostatistics, or biometry (must have received Ph.D. prior to start date with the Company) with graduate level course work, project, consulting or internship experience in: writing the statistics section a protocol or analyzing a clinical trial/case study, and working with a team to apply statistical methodology to a research question, and communicating basic statistical information to non-statisticians.
- Publishing: at least one publication (as primary or joint author) in a statistical, mathematical or clinical journal.
- Good oral and written communication skills.
- Attention to detail. Possess a strong quality orientation. Ensure tasks are completed correctly and on time.

Principal Biostatistician Requirements:

- Masters’ degree from an accredited institution required and 6 years’ experience within the pharmaceutical industry, CROs, regulatory authorities, or academic institutions, or Doctoral degree (PhD, MD) from an accredited institution with 3 years’ experience within the pharmaceutical industry, CROs, regulatory authorities, or academic institutions.
- Ability to interact with authorities on statistical issues at the trial level.
- Thorough knowledge of statistical methodology, processing clinical trial information and the drug development process.
- Ability to communicate statistical information to non-statisticians.
- Ability to write publications (as joint author) in clinical trials.
- Excellent oral and written communication skills.
- Ability to manage project from a statistical perspective.
- Demonstrated ability to design, conduct and analyze a complex trial.
- Evidence of strong teamwork in order to successfully work with a trial team and project level team members.
Sr. Principal Biostatistician Requirements:

- Ph.D. in statistics, biostatistics, or biometry; at least 6 years’ experience in pharmaceutical clinical trial experience, preferably in the pharmaceutical industry and/or Regulatory Authorities, or M.S. in the above mentioned areas with 10 years of similar experience.
- Ability to work on local/global project teams in order to come to resolution on a project.
- Ability to mentor, motivate, teach a scientific/technical staff.
- Offer scientific insight to projects.
- Excellent interpersonal skills with the ability to interact effectively with people, internally and externally at all levels of the organization.
- Must have exceptional oral and written presentation skills.

Sr. Principal Biostatistician Desired Experience, Skills and Abilities:

- Record of publications (principal author) in methodological research.
- Extensive knowledge of scientific area of responsibility; ability to ask critical scientific questions and to critique devised hypotheses, experimental designs and results interpretation.
- Demonstrated ability to successfully plan and conduct a statistical analysis on research.
- Demonstrated ability in supervising scientific/technical work.

Eligibility Requirements:

- Must be legally authorized to work in the United States without restriction.
- Must be willing to take a drug test and post-offer physical (if required)
- Must be 18 years of age or older

Our Culture:
Boehringer Ingelheim is one of the world’s top 20 pharmaceutical companies and operates globally with approximately 50,000 employees. Since our founding in 1885, the company has remained family-owned and today we are committed to creating value through innovation in three business areas including human pharmaceuticals, animal health and biopharmaceutical contract manufacturing. Since we are privately held, we have the ability to take an innovative, long-term view. Our focus is on scientific discoveries and the introduction of truly novel medicines that improve lives and provide valuable services and support to patients and their families. Employees are challenged to take initiative and achieve outstanding results. Ultimately, our culture and drive allows us to maintain one of the highest levels of excellence in our industry. We are also deeply committed to our communities and our employees create and engage in programs that strengthen the neighborhoods where we live and work. Boehringer Ingelheim, including Boehringer Ingelheim Pharmaceuticals, Inc., Boehringer Ingelheim USA, Boehringer Ingelheim Animal Health USA, Inc., Merial Barceloneta, LLC and Boehringer Ingelheim Fremont, Inc. is an equal opportunity and affirmative action employer committed to a culturally diverse workforce. All qualified applicants will receive consideration for employment without regard to race; color; creed; religion; national origin; age; ancestry; nationality; marital, domestic partnership or civil union status; sex; gender identity or expression; affectional or sexual orientation; disability; veteran or military status, including protected veteran status; domestic violence victim status; atypical cellular or blood trait; genetic information (including the refusal to submit to genetic testing) or any other characteristic protected by law.

Boehringer Ingelheim is firmly committed to ensuring a safe, healthy, productive and efficient work environment for our employees, partners and customers. As part of that commitment, Boehringer Ingelheim conducts pre-employment verifications and drug screenings.

Use the following link to submit your resume for consideration:

https://tas-boehringer.taleo.net/careersection/global+template+career+section+28external29/jobdetail.ftl?job=1811399&tz=GMT-05%3A00

32 2019 ICSA Applied Statistics Symposium, Raleigh, NC, June 9-12
## Statistics

The latest release of SAS/STAT® is now available. SAS/STAT 15.1 provides new methodology and new capabilities:

### SAS/STAT 15.1 Highlights

- Bayesian generalized linear mixed models.
- Graphical causal models.
- Regression for time-to-event data based on restricted mean survival time.
- Counterfactual analysis using quantile regression.
- Semiparametric proportional hazards model for interval-censored data.

### Recent SAS/STAT Additions

- Casual 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.

Learn more
support.sas.com/statnewreleases
Explore partnering with Sanofi Pasteur

Building on its history of successful collaborations, Sanofi Pasteur, the vaccines Global Business Unit of Sanofi, is seeking partners with a common drive for excellence and pursuit of innovation.

Sanofi Pasteur is interested in potential partnering opportunities in the field of active and passive human immunization for infectious diseases, as well as technologies supporting product development and industrial performance, including in the following areas:

- Vaccines, monoclonal antibodies and supporting technologies for prevention and treatment of infectious diseases
  - Novel antigens and methods for antigen discovery and characterization
  - Vaccine vectors suitable for nasal or oral use
  - New ways to administer vaccines
  - Carrier proteins and protein-polysaccharide conjugation methods or alternative technologies

- Agents to enhance vaccine immune responses
  - Adjuvants and immunomodulators
  - Vaccine vectors and delivery systems intended to enhance or modify immune responses
  - Biological and immunological studies to further characterize adjuvants and immunomodulators

- Characterization and assay of immune responses and disease markers
  - Animal models of human diseases
  - Biological markers for evaluating the efficacy of prophylactic or therapeutic interventions
  - In vitro models of human tissues, including the immune system
  - Epidemiological studies relevant to the use of vaccines and immunotherapeutics
  - 3D tissue models

- Tools for improving vaccine and monoclonal antibody research, development and production
  - Development and application of new technologies in the areas of genomics and proteomics
  - Prokaryotic or eukaryotic cell lines for antigen production
  - Fermentor and bioreactor technology
  - Disposable systems
  - Online testing
  - Downstream processing; purification and aseptic filling processes
  - Process automation
  - Preservatives and stabilizers
  - Bioinformatics techniques for modeling, data handling and analysis
  - Anti-counterfeiting technology

Sanofi Pasteur has a strong commitment to establishing research and development partnerships with major universities, research institutes, government bodies, biotechnology companies and contract research organizations. The company’s collaborations cover virtually all aspects of vaccine development, including early-stage research.

Examples of current partnerships and technology investments include a protective mAb against RSV infection for infants, vaccine candidates against Zika virus, RSV, HSV, S.pneumoniae, and broadly protective influenza; pediatric combination vaccines; large-scale cell-culture based virus production; adjuvants and immuno-modulators; conjugate vaccine production; and vaccine delivery systems.

A company partnering with Sanofi Pasteur interacts with a multidisciplinary team that has years of experience in working to ensure that partnerships are executed successfully and nurtured for the mutual benefit of all parties. This approach utilizes the value-added Sanofi Pasteur alliance management capability, which focuses on the relationship by facilitating open communication, trust, understanding and clear expectations across the project life span.

Combined with the technical competency of the alliance, this balance provides a well-rounded environment in which novel technologies can flourish. Currently, 100% of its preclinical portfolio and ~65% of its clinical portfolio has a partnering component.

Sanofi Pasteur welcomes information about new partnership opportunities. Each opportunity is carefully evaluated and reviewed by its dedicated team.

Contact:

Roman Chicz, Head, External Research and Development
Tel: +1 617 866 4562
Email: roman.chicz@sanofi.com

Jean-Louis Grunwald, Vice President, Business Development
Tel: +65 64 31 22 10
Email: jean-louis.grunwald@sanofi.com
Promoting statistical excellence and championing statistical influence, Pfizer Statisticians employ innovative study designs and best statistical practices for data analyses and interpretation.

Interested in pursuing a career in statistics at Pfizer?

http://www.pfizer.com/careers

Pfizer
Breakthroughs that change patients’ lives
Why We Invent

AT MERCK, WE ARE INVENTING FOR LIFE.

We are taking on many of the world’s most challenging diseases because the world still needs cures for cancer, Alzheimer’s disease, HIV, and so many other causes of widespread suffering in people and animals.

We invent to help people go on, unburdened, to experience, create and live their best lives.

Copyright © 2017 Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc.
All Rights Reserved. C01RP-104478-005 06/17
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.
You are cordially invited to the 2020 ICSA Applied Statistics Symposium in Houston!
The ICSA Applied Statistics Symposium will be held from Sunday, May 17 to Wednesday, May 20, 2020 at The Westin Galleria Houston, 5060 W Alabama Street, Houston, Texas. Please send any inquiry to Dr. Hulin Wu (Hulin.Wu@uth.tmc.edu). Houston is the fourth-most populous city in the United States. Houston is a global city, with strengths in business, international trade, entertainment, culture, media, fashion, science, sports, technology, education, medicine, and research. It is home to many cultural institutions and exhibits, which attract more than 7 million visitors a year. Notable facilities include the Museum of Fine Arts, Houston Museum of Natural Science, the Contemporary Arts Museum Houston, the Station Museum of Contemporary Art, Holocaust Museum Houston, the Houston Zoo, and NASA’s Lyndon B. Johnson Space Center. Houston is the seat of the internationally renowned Texas Medical Center (TMC), which contains the world’s largest concentration of research and healthcare institutions including University of Texas Health Science Center at Houston, MD Anderson Cancer Center, Baylor College of Medicine, Memorial Hermann Hospital, The Methodist Hospital, and Texas Children’s Hospital. Several universities of higher education such as The University of St. Thomas, University of Houston, and Rice University are located within the city. 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 1, 2019.

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 February 15, 2020.

Call for Short Course Proposals
Anyone who is interested in giving a one-day or half-day short course at 2020 ICSA Applied Statistics Symposium is welcome to submit a short-course proposal to Dr. Wenyi Wang (wwang7@mdanderson.org). The submission deadline is December 15, 2019.

Executive Committee
- Executive Committee Chair: Hulin Wu (University of Texas Health Science Center at Houston)
- Scientific Program Chair: Momiao Xiong (University of Texas Health Science Center at Houston) and Jianhua Huang (Texas A&M University)
- Program Book and Website Committee Chair: Yunxin Fu and Ashraf Yaseen (University of Texas Health Science Center at Houston)
- Local Committee Chair: Hongyu Miao (University of Texas Health Science Center at Houston)
- Treasurer: Dejian Lai (University of Texas Health Science Center at Houston)
- Student Paper Competition Committee Chair: Ruosha Li (University of Texas Health Science Center at Houston) and Jing Ning (University of Texas MD Anderson Cancer Center)
- Short Course Committee Chair: Wenyi Wang (University of Texas MD Anderson Cancer Center)
- Fund Raising Committee Chair: Rui (Sammi) Tang (Servier Pharmaceuticals)
Announcement: The 2020 ICSA China Conference
June 27-30, 2020, Chengdu, China

The 2020 ICSA China Conference will take place, June 27-30 2020, at Jinniu Hotel, Chengdu, China. 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 Chair: Professor Jianxin Pan, University of Manchester, jianxin.pan@manchester.ac.uk.
June 10 9:10-10:10

Keynote Session I (Keynote)
Room: RCC Ballroom C
Organizer: ICSA 2019 organizing committee.
Chair: Lanju Zhang, Abbvie, Inc.

9:10 Keynote lecture I: Leveraging Real-World Data for Regulatory Decision-Making: Propensity Score-Based Patient Selection and Augmentation in Investigational Clinical Studies
Lilly Yue. U.S. Food and Drug Administration

June 10 10:30 - 12:10

Session 1: Modern Methods for Neuroimaging Data (Invited)
Room: 301A
Organizer: Alexander McLain, University of South Carolina.
Chair: Feifei Xiao, University of South Carolina.

10:30 Tensor-on-tensor regression
Eric Lock. University of Minnesota

10:55 A Multivariate Generalized Linear Mixed Model for Joint Modeling of Cognitive and Neuroimaging Outcomes
*Mulugeta Gebregziabher1, Carter Allen1, Daniel Baer2, Aastha Khatiwada3, Virginia Shipes1, Abeba Teklehaimanot4 and Philip Insel. 1Medical University of South Carolina

11:20 Bayesian Spatial Binary Regression for Label Fusion in Structural Neuroimaging
*Andrew Brown1, Chris Memarian2, Russell Shinohara2 and Kristin Linn2. 1Clemson University 2University of Pennsylvania

11:45 Weighted p-values for heterogeneous dependent data: an application to lesion symptom mapping
*Alexander Mcclain1, Joshua Habiger2, Christopher Rorden1 and Julius Fridriksson1. 1University of South Carolina 2Oklahoma State University

12:10 Floor Discussion.

Session 2: Statistical Advances in Genetics, Genomics and Bioinformatics (Invited)
Room: 301B
Organizer: Zuoheng Wang, Yale University.
Chair: Weimiao Wu, Yale University.

10:30 A unified method for rare variant analysis of gene-environment interactions
Elise Lim1, Han Chen2, Josee Dupuis3 and *Ching-Ti Liu1. 1Boston University 2University of Texas Health Science Center at Houston

10:55 Stochastic Functional Linear Models for Gene-based Association Analysis of Quantitative Traits
*Razong Fan1, Shuqi Wang1, Xiaohan Mei3, Bingsong Zhang1, Yue Han1, Runqiu Wang1 and Fang Hong-Bin. 1Georgetown University Medical Center

11:20 Gene-graph based imputation method for single-cell RNA sequencing data
*Weimiao Wu, Yuning Liu, Geoffrey Chupp, Xiting Yan and Zuoheng Wang. Yale University

11:45 Identification of trans-eQTLs using mediation analysis with multiple mediators
Nayang Shan1, Zuoheng Wang2 and Lin Hou1. 1Tsinghua University 2Yale University

12:10 Floor Discussion.

Session 3: Recent development in missing data and survey sampling problems (Invited)
Room: 302A
Organizer: Sixia Chen, University of Oklahoma Health Sciences Center.
Chair: Sixia Chen, University of Oklahoma Health Sciences Center.

10:30 A Bayesian Growth Mixture Model for Complex Survey Data: Clustering Post-Disaster PTSD Trajectories
Rebecca Anthopolos1, Qixuan Chen1, Joseph Sedransk2, Mary Thompson2, Gang Meng3 and Sandro Galea4. 1Columbia University 2Joint Program in Survey Methodology, University of Maryland 3University of Waterloo 4Boston University

10:55 Targeting key survey variables at the nonresponse treatment stage
*David Haziza1, Sixia Chen2 and Yineng Gao. 1Université de Montréal 2University of Oklahoma

11:20 Improving Population Representativeness for Estimating Prevalence of Risk-factors and Diseases
*Lingxiao Wang1, Barry Graubard2, Hormuzd Katki2 and Yan Li3. 1University of Maryland, College Park 2Biostatistics Branch, National Cancer Institute

11:45 Measures of the Degree of Departure from Ignorable Sample Selection for Non-Probability Samples
*Brady T. West1, Roderick Little1, Phil Boonstra1 and Rebecca Andridge2. 1University of Michigan 2Ohio State University

12:10 Floor Discussion.

Session 4: Advances in statistical methods for large-scale genomic studies (Invited)
Room: 302B
Organizer: Hui Jiang, University of Michigan.
Chair: Hui Jiang, University of Michigan.

10:30 RNA-seq and Methyl-seq power calculation and design issue
Chien-Wei Lin1 and *George Tseng2. 1Medical College of Wisconsin 2University of Pittsburgh
Session 5: Analyzing big and complex data using modern machine learning techniques (Invited)
Room: 302C
Organizer: Jiwei Zhao, State University of New York at Buffalo.
Chair: Jiwei Zhao, State University of New York at Buffalo.
10:30 Double-Slicing Assisted Sufficient Dimension Reduction for High Dimensional Censored Data
Shanshan Ding\textsuperscript{1}, Wei Qian\textsuperscript{2} and Lan Wang\textsuperscript{2}.\textsuperscript{1}University of Delaware\textsuperscript{2}University of Minnesota
10:55 Optimal tuning for divide-and-conquer kernel ridge regression with massive data
Ganggang Xu\textsuperscript{1}, ZuoFeng Shang\textsuperscript{2} and Guang Cheng\textsuperscript{3}.\textsuperscript{1}University of Miami\textsuperscript{2}Indiana University-Purdue University Indianapolis\textsuperscript{3}Purdue University
11:20 Integrative Linear Discriminant Analysis
Quefeng Li\textsuperscript{1} and Lexin Li\textsuperscript{2}.\textsuperscript{1}University of North Carolina at Chapel Hill\textsuperscript{2}University of California, Berkeley
11:45 Bayesian Deep Neural Networks for Nonlinear Variable Selection
Yan Sun, Qifan Song and Faming Liang. Purdue University
12:10 Floor Discussion.

Session 6: Recent advances in statistical network modeling: go beyond simple network data (Invited)
Room: 305A
Organizer: Tianxi Li, University of Virginia.
Chair: Tianxi Li, University of Virginia.
10:30 Hierarchical community detection by recursive partitioning
Tianxi Li\textsuperscript{1}, Liu Hu\textsuperscript{2}, Sharmodeep Bhattacharyya\textsuperscript{3}, Purnamrita Sarkar\textsuperscript{4}, Peter Bickel\textsuperscript{5} and Elizaveta Levina\textsuperscript{5}.\textsuperscript{1}University of Virginia\textsuperscript{2}University of California, Berkeley\textsuperscript{3}Oregon State University\textsuperscript{4}University of Texas at Austin\textsuperscript{5}University of Michigan
10:55 Finding the way for the Graph Matching Problem
Daniel Sussman\textsuperscript{1} and Vince Lyzinski\textsuperscript{2}.\textsuperscript{1}Boston University\textsuperscript{2}University of Massachusetts Amherst
11:20 Transform-based unsupervised point registration and unseeded low-rank graph matching
Yuan Zhang. Ohio State University
11:45 Floor Discussion.

Session 7: Recent Advances in Empirical Likelihood Methods (Invited)
Room: 305B
Organizer: Fei Gao, University of Washington.
Chair: Qingning Zhou, University of North Carolina at Charlotte.
10:30 Empirical Likelihood-based Inferences for Median Medical Cost Regression Models with Censored Data
Gengsheng Qin and Guanhao Wei. Gerogia State University
10:55 Empirical likelihood with two samples for non-separable parameters
Xie Ding and Mai Zhou. University of Kentucky
11:20 A New Scope of Penalized Empirical Likelihood in High Dimensionality
Tong Tong Wu\textsuperscript{1}, Chengyong Tang\textsuperscript{2} and Jinyuan Chang\textsuperscript{3}.\textsuperscript{1}University of Rochester\textsuperscript{2}Temple University\textsuperscript{3}Southwestern University of Finance and Economics
11:45 Non-iterative Estimation for Semiparametric Models with Population-based Auxiliary Information
Fei Gao and Kwun Chuen Gary Chan. University of Washington
12:10 Floor Discussion.

Session 8: Functional data analysis: Foundations, Tools and Applications (Invited)
Room: 306A
Organizer: Luo Xiao, North Carolina State University
Chair: Raymond K. W. Wong, Texas A&M University.
10:30 Modeling Longitudinal Data on Riemannian Manifolds
Xiongtao Dai\textsuperscript{1}, Zhenhua Lin\textsuperscript{2} and Hans-Georg Mueller\textsuperscript{2}.\textsuperscript{1}Iowa State University\textsuperscript{2}University of California, Davis
10:55 tidyfun: a New Framework for Representing and Working with Function-Valued Data
Fabian Scheipl\textsuperscript{1} and Jeff Goldsmith\textsuperscript{2}.\textsuperscript{1}Ludwig-Maximilians-Universität München\textsuperscript{2}Columbia University
11:20 Finding Biomarkers for Childhood Obesity Using Functional Data Analysis
Sarah Craig, Ana Kenney, Francesca Chiaromonte, Matthew Reimherr and Kateryna Makova. Pennsylvania State University
11:45 A Functional Mixed Model for Scalar on Function Regression with Application to Functional MRI Study
Wanying Ma\textsuperscript{1}, Luo Xiao\textsuperscript{1}, Bowen Liu\textsuperscript{1} and Martin Lindquist\textsuperscript{2}.\textsuperscript{1}North Carolina State University\textsuperscript{2}Johns Hopkins University
12:10 Floor Discussion.

Session 9: Innovative methods for complex survival data and the applications (Invited)
Room: 306B
Organizer: Song Yang, National Institutes of Health
Chair: Colin Wu, National Institutes of Health.
Session 10: Classification and modeling of multiple time series and functional data and applications (Invited)
Room: 306C
Organizer: Ming-Yen Cheng, Hong Kong Baptist University.
Chair: Ming-Yen Cheng, Hong Kong Baptist University.

10:30 Development and Evaluation of Sensor Fire Prediction for Smart Appliances
Z. Q. John Lu, Amy Mensch and Anthony Hammins. National Institute of Standards and Technology

10:55 Evaluating classification accuracy for modern learning approaches
Jialiang Li. National University of Singapore

11:20 A multi-group classification model for geographical origin discrimination of herbal medicines
Ying Zhu and Tuck Lee Tan. Nanyang Technological University

11:45 A simple and adaptive two-sample test in high dimensions
Jin-Ting Zhang¹, Jia Guo², Bu Zhou³ and Ming-Yen Cheng⁴. ¹National University of Singapore ²Zhejiang University of Technology ³Zhejiang Gongshang University ⁴Hong Kong Baptist University

12:10 Floor Discussion.

Session 11: Statistical Advances in High Dimensional Data Analysis (Invited)
Room: 303
Organizer: Feifei Xiao, University of South Carolina.
Chair: Jiwei Zhao, University at Buffalo.

10:30 Co-Manifold Learning with Missing Data
Eric Chi¹, Gal Mishne² and Ronald Coifman². ¹North Carolina State University ²Yale University

10:55 Identifying noncoding RNAs and methylation biomarkers associated with tumor immunity
Xuefeng Wang and Xiaoping Yu. H. Lee Moffitt Cancer Center & Research Institute

11:20 LDSaRa: A Powerful Method for High Resolution Copy Number Variation Detection
Feifei Xiao and Xizhi Luo. University of South Carolina

11:45 A Bayesian sparse latent factor model for identification of cancer subgroups with data Integration
Zequn Sun, Brian Neelon and Dongjun Chung. Medical University of South Carolina

12:10 Floor Discussion.

Session 12: Challenges in implementation of ICH E17 (Invited Panel)
Room: 304
Organizer: Nai-te Ting, Boehringer Ingelheim, Inc.
Chair: Songqiao Huang, Boehringer Ingelheim, Inc.

10:30 Experience of Reviewing MRCT Results and Expectations for ICH E17 in Taiwan
Guei-Feng Tsai. Statistical Team, CDE Taiwan

10:55 Panelists:
James Hung, US Food and Drug Administration
William Wang, Merck & Co.

11:20 Discussant: Mary Zhao, Boehringer Ingelheim, Inc

11:55 Floor Discussion.

Session 13: Historical Control in Clinical Trials (Invited)
Room: 307
Organizer: rui (sammi) tang, Shire.
Chair: rui (sammi) tang, shire.

10:30 The Use of Historical Control in Clinical Trials in Rare Diseases
Chungin Deng, Youlan Rao and Lisa Edwards. United Therapeutics

10:55 Statistical methodologies and challenges to properly unitize historical data
Rong Liu¹ and Mercedes Ghadessi². ¹Celgene ²Bayer Co

11:20 Actual example of using historical control data for drug development
Kiichiro Toyoizumi. Shionogi, Inc.

11:45 Discussant: Robert Beckman, Georgetown University Medical Center

12:10 Floor Discussion.

Session 14: Using adaptive treatment strategies to give the right patient the right dose at the right time (Invited)
Room: 301A
Organizer: Tianyu Zhan, Abbvie, Inc.
Chair: Alan Hartford, Abbvie, Inc.

14:00 Optimal Individualized Treatment Initiation Regime
Xin Chen¹, Rui Song², Jiajia Zhang³, Liusuan Sun¹ and Wenbin Lu². ¹Chinese Academy of Science ²North Carolina State University ³University of South Carolina

14:25 Subgroup-Specific Dose-Finding in Phase I Trials Based on Time to Toxicity
Peter Thall¹ and Andrew Chapple². ¹M.D. Anderson Cancer Center ²Louisiana State University
14:50 Improving Trial-based Treatment Strategies Using Evidence from Electronic Health Records
Peng Wu1, Donglin Zeng2, Haoda Fu3 and Yuanjia Wang4.
1Columbia University 2University of North Carolina at Chapel Hill 3Eli Lilly and Company
15:15 Floor Discussion.

Session 15: Breakthroughs in the Analysis of Large-Scale and Complex Data (Invited)
Room: 301B
Organizer: Jessie Jeng, North Carolina State University.
Chair: John Daye, Independent researcher.
14:00 Efficient Signal Inclusion in Large-Scale Data Analysis
Jessie Jeng. North Carolina State University
14:25 Dynamic Scan Procedure for Detecting Rare-Variant Association Regions in Whole Genome Sequencing
Xihong Lin. Harvard School of Public Health
14:50 Analysis of clustered data with cluster-specific nonignorable nonresponse
Peisong Han. University of Michigan
15:15 Two component semiparametric density mixture with a known component: estimation and applications
Zhou Shen1 and Michael Levine2. 1J.P. Morgan 2Purdue University
15:40 Floor Discussion.

Session 16: New Statistical Methods for Missing Data and Biased Sampling (Invited)
Room: 302A
Organizer: Qingning Zhou, University of North Carolina at Charlotte.
Chair: Fei Gao, University of Washington.
14:00 Propensity score calibration for missing-at-random data
Peisong Han. University of Michigan
14:25 Semiparametric inference for merged data from multiple sources
Takumi Saegusa. University of Maryland, College Park
14:50 Variable selection for multiple types of high-dimensional features with missing data
Kin Yau Wong1, Donglin Zeng2 and Danyu Lin2. 1The Hong Kong Polytechnic University 2University of North Carolina at Chapel Hill
15:15 Case-cohort Studies with Multiple Interval-censored Disease Outcomes
Qingning Zhou1, Jianwen Cai2 and Haibo Zhou2. 1University of North Carolina at Charlotte 2University of North Carolina at Chapel Hill
15:40 Floor Discussion.

Session 17: Different quantities of interest in observational studies (Invited)
Room: 302B
Organizer: Peisong Han, University of Michigan.
Chair: Peisong Han, University of Michigan.
14:00 Statistical Inference on Shape and Size Indexes for Recurrent Event Processes
Yifei Sun1, Steven Chiou2 and Chiung-Yu Huang3. 1Columbia University 2University of Texas at Dallas 3University of California, San Francisco
14:25 Parsimonious Regressions for Repeated Measure Analysis
Dennis Cook1, Liliana Forzani and Lan Liu1. 1University of Minnesota
14:50 When Wilcoxon-Mann-Whitney Meets Non-Compliance
La Mao. University of Wisconsin-Madison
15:15 Solar Flare Predictions with Machine Learning
15:40 Floor Discussion.

Session 18: Quantitative decision making in early clinical development (Invited)
Room: 302C
Organizer: Weidong Zhang, Pfizer, Inc..
Chair: Weidong Zhang, Pfizer, Inc..
14:00 An Evidence Based Approach for Phase II Proof-of-Concept Trial
Satrajit Roychoudhury. Pfizer, Inc.
14:25 Considerations for Basket Trial Design under Multisource Exchangeability Assumptions
Michael Kane1, Alex Kaizer2, Nan Chen3 and Brian Hobbs1. 1Yale University 2University of Colorado Denver 3University of Texas MD Anderson Cancer Center 4The Cleveland Clinic
14:50 Quantitative decision making in proof of mechanism study
Weidong Zhang. Pfizer, Inc.
15:15 Discussant: Naitee Ting, Boehringer Ingelheim, Inc
15:40 Floor Discussion.

Session 19: Causal inference and missing data analysis: identification and estimation (Invited)
Room: 305A
Organizer: Shu Yang, North Carolina State University.
Chair: Sixia Chen, University of Oklahoma.
14:00 Causal inference with confounders missing not at random
Shu Yang1, Linbo Wang2 and Peng Ding3. 1North Carolina State University 2University of Toronto 3University of California, Berkeley
14:25 The Blessings of Multiple Causes
Yixin Wang and David Blei. Columbia University
14:50 Analysis of clustered data with cluster-specific nonignorable nonresponse
Danhyang Lee and Jae-Kwang Kim. Iowa State University
15:15 Recent Advances and Best Practices in Comparative Analyses using Real World Data
Douglas Faries. Eli Lilly and Company
15:40 Floor Discussion.
Session 20: Design and analysis of complex survival data (Invited)
Room: 305B
Organizer: Qing Yang, Duke University.
Chair: Qing Yang, Duke University.
14:00 Joint modeling of survival and longitudinal quality of life data with informative censoring
Zhigang Li1, Yanaka Peragaswaththiyan1 and Lihui Zhao2. 1University of Florida 2Northwestern University
14:25 Sample size calculation for cluster randomization trials with a time-to-event endpoint
• Sim-Ho Jung1 and Jianghao Li. 1Duke University
14:50 Sample Size Calculation for Studies with Grouped Survival Data
Zhiguo Li, Xiaofei Wang, Yuan Wu and Kouros Owzar. Duke University
15:15 SampleN: an R-shiny APP to calculate sample size, including subgroup analysis for survival outcomes
Qian Wu and Isaac Jenkins. Fred Hutchinson Cancer Research Center
15:40 Floor Discussion.

Session 21: Nonconvex Optimization and Panel Data Analysis (Invited)
Room: 306A
Organizer: Quefeng Li, University of North Carolina at Chapel Hill.
Chair: Quefeng Li, University of North Carolina at Chapel Hill.
14:00 Nonconvex Regularized Robust Regression with Oracle Properties in Polynomial Time
• Wensh Zhou1, Xiaou Pan2 and Qiang Sun2. 1University of Carolina, San Diego 2University of Toronto
14:25 On the Global Convergence of Policy Optimization in Deep Reinforcement Learning
Zhaoan Wang1, Qi Cai, Jason Lee, Boyi Liu and Zhuoran Yang. 1Northwestern University
14:50 Structure Learning in Panel Data Analysis
Yuan Ke1, Jiatiang Li2 and Wenyang Zhang3. 1University of Georgia 2National University of Singapore 3University of York
15:15 Floor Discussion.

Session 22: Recent advances on high dimensional statistical learning and algorithm (Invited)
Room: 306B
Organizer: Yufeng Liu, University of North Carolina at Chapel Hill.
Chair: Yufeng Liu, University of North Carolina at Chapel Hill.
14:00 First-order Newton-type Estimator for Distributed Estimation and Inference
• Xi Chen1, Weidong Liu2 and Yichen Zhang1. 1New York University 2Shanghai Jiao Tong University
14:25 Joint association and classification analysis of multi-view data
Yunfeng Zhang and • Irina Gaynanova. Texas A&M University
14:50 Ensemble estimation and variable selection with semiparametric regression models
• Sunyoung Shin1, Yufeng Liu2, Stephen Cole2 and Jason Fine2. 1University of Texas at Dallas 2University of North Carolina at Chapel Hill
15:15 Recovery of Sums of Sparse and Dense Signals by Incorporating Graphical Structure Among Predictors
• Yiyun Luo and Yufeng Liu. University of North Carolina at Chapel Hill
15:40 Floor Discussion.

Session 23: Use of Real-World Data and Real-World Evidence in the Regulatory Decision-making of Medical Products (Invited)
Room: 306C
Organizer: Nelson Lu, Center for Devices and Radiological Health, FDA.
Chair: Qin Li, Center for Devices and Radiological Health, FDA.
14:00 Collaborative Effort from to use RWD Establish Performance Criteria for medical devices.
• Roseann White1, Yu-Ching Cheng2, Tianyi Sun3, Niveditta Ramakumar4 and Daniel Berges5. 1Your Third Opinion, 2U.S. Food and Drug Administration, 3Weill School of Medicine, Cornell University 4Geisel School of Medicine, Dartmouth University 5University of Vermont Medical Center
14:25 Making a comparative claim when leveraging information of the control formed from real-world data
• Nelson Lu, Wei-Chen Chen and Yanling Xu. Center for Devices and Radiological Health, FDA
14:50 Incorporating Real-World Evidence in Single-Arm Clinical Studies
• Chenguang Wang1, Heng Li2, Wei-Chen Chen2, Nelson Lu2, Ram Tiwari2, Yanling Xu2 and Lilly Yue. 1Johns Hopkins University 2U.S. Food and Drug Administration, CDRH.
15:15 Synthesizing Real World Data into Evidence for Medical Devices
Gregory Campbell. GCStat Consulting, LLC
15:40 Floor Discussion.

Session 24: Master Protocols: Benefits, Opportunities and Challenges (Invited Panel)
Room: 303
Organizer: Zoran Antonijevic, Z-Adaptive Design.
Chair: Zoran Antonijevic, Z-Adaptive Design.
14:00 Panelists:
Robert Beckman, Georgetown University Medical Center
Scott Berry, Berry Consultants, LLC.
Kun He, US Food and Drug Administration
Zhaoing Meng, Gates Medical Research Institute
Richard Simon, R Simon Consulting
14:25 Floor Discussion.
Session 25: Statistical and General Considerations on Development of Gene Therapies (Invited)
Room: 304
Organizer: Yang Song, Vertex Pharmaceuticals, Inc.
Chair: Yang Song, Vertex Pharmaceuticals, Inc.
14:00 Statistical challenges in the development of gene therapies for rare genetic diseases
Wenliang Shi, Bluebird Bio, Inc.
14:25 Regulatory Experiences in the Development and Approval of Gene Therapies
Showyuen Lee, U.S. Food and Drug Administration, CBER.
14:50 Estimand and patient journey - Analysis considerations for CAR-T Phase 3 studies
Yiyun Zhang, Novartis
15:15 Design of Clinical Study for Rare Diseases via Borrowing Historical Data
15:40 Floor Discussion.

Session 26: Recent Advances in High-dimensional Data and Nonparametric Statistics (Invited)
Room: 307
Organizer: Guanqun Cao, Auburn University.
Chair: Guanqun Cao, Auburn University.
14:00 Restricted function-on-function regression models
Xin Qi and Ruiyan Luo, Georgia State University
14:25 Functional regression for highly densely observed data with novel regularization
Ruiyan Luo and Xin Qi, Georgia State University
14:50 Estimation and Inference for Generalized Geoadditive Models
Shan Yu, Guannan Wang, Li Wang, Chenhui Liu and Lijian Yang, Iowa State University
15:15 Distance-Based Analysis with Quantile Regression Models
Shaoyu Li, Yanqing Sun, Liyang Diao and Xue Wang, University of North Carolina at Charlotte.
15:40 Floor Discussion.

June 10 16:00 - 17:40

Session 27: Statistical Learning for Complex Data (Invited)
Room: 301A
Organizer: Linglong Kong, University of Alberta.
Chair: Ruoning Zhu, University of Illinois at Urbana-Champaign.
16:00 High-dimensional Vector Autoregressive Modeling via Tensor Decomposition
Di Wang, Heng Lian, Wei Keung Li and Guodong Li. University of Hong Kong
Yeying Zhu, Donna Coffman and Debashis Ghosh. University of Waterloo
16:50 Ensembling Imbalanced-Spatial-Structured Support Vector Machine
Wenqing He, Xin Liu and Grace Yi, University of Western Ontario
17:15 Symmetrization for exact nonparametric functional ANOVA
Adam B Kashlak, University of Alberta
17:40 Floor Discussion.

Session 28: Statistical Methods for Feature Explorations in Omics Data Analysis (Invited)
Room: 301B
Organizer: Jung-Ying Tzeng, North Carolina State University.
Chair: Yi-Ju Li, Duke University.
16:00 Dynamic correlation analysis and its applications in RNA-seq Data
Tianwei Yu, Emory University
16:25 Whole Genomes Classification and Clustering Based on Nucleotide Positions Distributions
Hsin-Hsiung Huang, University of Central Florida
16:50 Estimation of Insulin Resistance using Existing Untargeted Metabolomics Profiling
Fang-Chi Hsu, NichOLEtTE Palmer, ShyH-Huei Chen and Donald Bowden, Wake Forest University
17:15 p-value Adjustment Procedure Using Empirical Weights
Yen-Yi Ho, Chong Ma, Wenda Zhang, Stephanie Christenson and Richard Nho, University of South Carolina
17:40 Floor Discussion.

Session 29: Statistical Machine Learning Methods for Diagnostic Medicine and Biomarkers (Invited)
Room: 302A
Organizer: Qiang Sun, University of Toronto
Chair: Jiali Li, National University of Singapore.
16:00 A New Information Criterion for Optimal Model Selection
Jie Ding, Vahid Tarokh and Yuhong Yang, University of Minnesota
16:25 Biomarker models for early Alzheimers disease diagnosis before symptoms appear
Zheyu Wang, Johns Hopkins University
16:50 Methods for predicting ovarian cancer diagnosis with a longitudinal biomarker CA-125
Yongli Han, Ming Wang and Danping Liu, National Cancer Institute
17:15 Floor Discussion.
Session 30: New Frontier in High-dimensional Data Analysis
(Invited)
Room: 302B
Organizer: Lan Wang, University of Minnesota.
Chair: Jiwei Zhao, State University of New York at Buffalo.
16:00 Nonregular and Minimax Estimation of Individualized Thresholds in High Dimension with Binary Responses
Huijie Feng¹, Yang Ning¹ and Jiwei Zhao². ¹Cornell University ²State University of New York at Buffalo
16:25 High-Dimensional Robust Thresholded Regression with Application to Scalar-on-Image Analysis
Lingzhou Xue. Pennsylvania State University
16:50 BET on Independence
Kai Zhang. University of North Carolina at Chapel Hill
17:15 OPTIMAL ADAPTIVITY OF SIGNED-POLYGON STATISTICS FOR NETWORK TESTING
Tracy Ke. Harvard University
17:40 Floor Discussion.

Session 31: Extreme value statistics (Invited)
Room: 302C
Organizer: Chen Zhou, Erasmus University Rotterdam.
Chair: Chen Zhou, Erasmus University Rotterdam.
16:00 Towards inference for multivariate time series extremes: multiple block sizes and overlapping blocks
Axel Bucher¹, Stanislav Volgushev² and Nan Zou². ¹Heinrich-Heine-Universitat Dusseldorf ²University of Toronto
16:25 Dynamic Multivariate Peak over Threshold Model
Zifeng Zhao. University of Notre Dame
16:50 Trends in extreme value indices
Laurens Dehaan and Chen Zhou. Erasmus University Rotterdam
17:15 Floor Discussion.

Session 32: Longitudinal Data Analysis with Finite Mixtures
(Invited)
Room: 305A
Organizer: Timothy OBrien, Loyola University Chicago.
Chair: Timothy OBrien, Loyola University Chicago.
16:00 On the improved estimation of the normal mixture components for longitudinal data
Tapio Nummi. Tampere University
16:25 Survival analysis with finite mixtures
Janne Salonen¹ and Jyrki Möttönen². ¹Finnish Centre for Pensions ²University of Helsinki
16:50 Practical Modelling and Design Aspects of Nonlinear Trajectory Analysis
Timothy OBrien and Molly Michalak. Loyola University Chicago
17:15 Floor Discussion.

Session 33: Recent advances in modern survival analysis and the novel applications (Invited)
Room: 305B
Organizer: Yichuan Zhao, Georgia State University.
Chair: Din Chen, University of North Carolina at Chapel Hill.
16:00 Multiplicative Rates Model for Recurrent Events in Case-Cohort Studies
Jianwen Cai¹, Poulami Maitra¹ and Leila Daf Amorim². ¹University of North Carolina at Chapel Hill ²Federal University of Bahia
16:25 Personalized Treatment Selection for Joint Optimization of Survival and Other Outcomes
Somnath Datta. University of Florida
16:50 Simultaneous Estimation and Variable Selection for Interval-censored Failure Time data
(Jony) Jianguo Sun. University of Missouri
17:15 Floor Discussion.

Session 34: Statistical Issues on Developing Medicines for Rare Diseases or Vulnerable Populations (Invited)
Room: 306A
Organizer: Shuyen Ho, UCB Biosciences, Inc..
Chair: Shuyen Ho, UCB Biosciences, Inc..
16:00 The Use of Historical Data in Assessing the Efficacy of Drugs in Rare Diseases
Ros Walley. UCB Pharma, Inc.
16:25 Design of a Phase 3 Trial for an Acute Treatment of a Rare Disease with Epidemic Attacks
Sharon Murray. BioCryst Pharmaceuticals, Inc.
16:50 Challenges, Achievements and Lessons Learned from two Clinical Studies in Women of Child Bearing Age
Maggie Wang. UCB Biosciences, Inc.
17:15 Floor Discussion.

Session 35: New frontiers in change point and complex data analyses (Invited)
Room: 306B
Organizer: Hanxiang Peng, Indiana University-Purdue University Indianapolis
Yichuan Zhao, Georgia State University.
Chair: Dongliang Wang, SUNY Upstate Medical University.
16:00 Empirical Likelihood for Change Point Detection in Autoregressive Models
Ramadha Piyadigamage¹ and Wei Ning². ¹Western Washington University ²Bowling Green State University
16:25 Robust Graph Change-point Detection for Brain Evolvement Study
Fang Han¹, Xi Chen², Honglang Wang³, Lexin Li³ and Brain Caffo³. ¹University of Washington ²New York University ³University of California, Berkeley ⁴Johns Hopkins University
16:50 Simultaneous Estimation and Variable Selection for Interval-censored Failure Time data
(Tony) Jianguo Sun. University of Missouri
17:15 Floor Discussion.
16:50 Testing Community Structure for Hypergraphs

- Mingao Yuan
- Ruiqi Liu
- Yang Feng
- Zuofeng Shang

1. North Dakota State University
2. Indiana University-Purdue University Indianapolis
3. Columbia University

17:15 Design based incomplete U-statistics

Xiangshun Kong

and

Wei Zheng

1. Beijing Institute of Technology
2. University of Tennesse

17:40 Floor Discussion.

**Session 36: Statistical Inference for Multi-layer data with Complex Dependence (Invited)**

Room: 306C

Organizer: Yumou Qiu, Iowa State University.
Chair: Yumou Qiu, Iowa State University.

16:00 Inference on Multi-level Partial Correlations based on Multi-subject Time Series Data

Yumou Qiu

Iowa State University

16:25 Hierarchical Functional Data Analysis

Yuhang Xu

University of Nebraska-Lincoln

16:50 Yield forecasting based on short time series with high spatial resolution data

Sayli Pokal, Yuzhen Zhou and Trenton Franz

University of Nebraska-Lincoln

17:15 Floor Discussion.

**Session 37: Modern Perspectives and Recent Advances in Quantile Regression (Invited)**

Room: 303

Organizer: Meng Li, Rice University.
Chair: Meng Li, Rice University.

16:00 Quantile regression of recurrent events risk

Huijuan Ma, Limin Peng, Chiueng-Ya Huang and Haoda Fu

1. Emory University

16:25 Counterfactual Analysis Using Quantile Regression

Yonggang Yao

SAS Institute, Inc.

16:50 Model Based Joint Quantile Regression with Dependency

Surya Tokdar

Duke University

17:15 Distributed Inference for Quantile Regression Processes

Stanislav Volgushev

1. University of Toronto
2. University of Missouri
3. Purdue University

17:40 Floor Discussion.

**Session 38: Object Oriented Data Analysis (Invited)**

Room: 304

Organizer: J. Steve Marron, University of North Carolina at Chapel Hill.
Chair: J. Steve Marron, University of North Carolina at Chapel Hill.

16:00 Local Hypothesis Testing for Functional Data: Extending FWER and FDR to the Functional Framework

Niels Asken Landtorp Olsen, Alessia Pini and Simone Vantini

1. University of Copenhagen
2. Universita Cattolica del Sacro Cuore (Italy)
3. Politecnico di Milano

16:25 PCA and Asymptotics on the Torus with Application to Biomolecules

Benjamin Eltzner, Stephan Huckemann and Kanti Mardia

1. University of Goettingen
2. University of Leeds
3. University of Oxford

16:50 Analysis of Populations of Networks: Structure Spaces and the Computation of Summary Statistics.

Anna Calissano, Aasa Feragen and Simone Vantini

1. Politecnico di Milano
2. University of Copenhagen

17:15 Discussant: J. Steve Marron, University of North Carolina at Chapel Hill

17:40 Floor Discussion.

**Session 39: Student Paper Award Session (Student)**

Room: 307

Organizer: Luo Xiao, North Carolina State University.
Chair: Luo Xiao, North Carolina State University.

16:00 Geostatistical Modeling of Positive Definite Matrices Using the Spatial Wishart Process

Zhou Lan, Brian J Reich, Joseph Guinness and Dipankar Bandyopadhyay

1. North Carolina State University
2. Cornell University
3. Virginia Commonwealth University

16:20 Integrative Analysis of Irregularly Measured and Mixed-Type Biomarkers in Electronic Health Records

Jitong Lou, Yuanjia Wang, Lang Li and Donglin Zeng

1. University of North Carolina at Chapel Hill
2. Columbia University
3. Ohio State University

16:40 GWAS-based AMD Progression Using a Copula Semiparametric Model

Tao Sun, Wei Chen and Ying Ding

University of Pittsburgh

17:00 Semiparametric Modelling and Estimation of the Global Percentile Outcome

Xiangyu Liu, Jing Ning, Xuming He and Ruoshua Li

1. University of Texas Health Science Center at Houston
2. University of Texas MD Anderson Cancer Center
3. University of Michigan

17:20 BART with Targeted Smoothing: an Analysis of Patient-Specific Stillbirth Risk

Jennifer Starling

University of Texas at Austin

17:40 Floor Discussion.

**June 11 8:00-9:00**

**Keynote Session II (Keynote)**

Room: RCC Ballroom C

Organizer: ICSA 2019 organizing committee.
Chair: Haoda Fu, Eli Lilly and Company.

8:00 Keynote lecture II: Are You Sure? The Power of Statistical Thinking

Steve Ruberg

Analytix Thinking, LLC


**June 11 9:10-10:10**

**Keynote Session III (Keynote)**
Room: RCC Ballroom C
Organizer: ICSA 2019 organizing committee.
Chair: Wenbin Lu, North Carolina State University.

9:10 Keynote lecture III: Modernizing Statistics Through Treat-ment Regimes: A Review
*Marie Davidian.* North Carolina State University

**June 11 10:30 - 12:10**

**Session 40: From bulk tissue to single cells: advances in statistical genetics and genomics (Invited)**
Room: 301A
Organizer: Yuchao Jiang, University of North Carolina at Chapel Hill.
Chair: Yang Chen, University of Michigan.

10:30 Fast and accurate alignment of single-cell RNA-seq samples using kernel density matching
*Yujie Chen, Qi Zhan* and *Yang Li.* University of Chicago

10:55 SCOPE: A normalization and copy number estimation method for single-cell DNA sequencing
*Rujin Wang, Daniyu Lin* and *Yuchao Jiang.* University of North Carolina at Chapel Hill

11:20 Benchmarking scRNA-seq clustering methods using multi-parameter ensembles of simulated data
*Xianing Zheng* and *Jun Li.* University of Michigan

11:45 Novel Deconvolution of Bulk Transcriptomics Data
*Fei Zou,* *Dong Li* and *Xiaojing Zheng.* University of North Carolina at Chapel Hill

12:10 Floor Discussion.

**Session 41: Advances in Meta-Analysis (Invited)**
Room: 301B
Organizer: Din Chen, University of North Carolina at Chapel Hill.
Chair: Din Chen, University of North Carolina at Chapel Hill.

10:30 Meta analysis of incidence of rare events using individual participant data
*Chen Chen,* *Yan Ma,* *Yong Ma* and *Qing Pan.* University of North Carolina at Chapel Hill

10:55 A Bayesian Hierarchical Model Estimating CACE in Meta-analysis of Randomized Clinical Trials
*Jincheng Zhou,* *James Hodges,* *M. Fareed Kuri* and *Haitao Chu.* University of Minnesota

11:20 On the efficiency of network meta-analysis
*Lifeng Lin.* Florida State University

11:45 Questionnaire on Network Meta-Analysis to Assess Its Rel-evance and Credibility
*Joseph Cappelleri.* Pfizer, Inc.

12:10 Floor Discussion.

**Session 42: Advance in Statistical Methods for Large and Complex Data (Invited)**
Room: 302A
Organizer: Dehan Kong, University of Toronto.
Chair: Yuying Xie, Michigan State University.

10:30 Comparing and weighting imperfect models using D-probabilities
*Yan Ma,* *Yong Ma* and *Qing Pan.* University of North Carolina State University

10:55 A Sparse Random Projection-based Test for Overall Qualitative Treatment Effects
*Chengchun Shi,* *Wenbin Lu* and *Rui Song.* North Carolina State University

11:20 Estimating the causal effect of treatment regimes for organ transplantation
*David Vock* and *Jeffrey Boatman.* University of Minnesota

11:45 A general Classification framework to estimate the optimal dynamic treatment regime
*Baoyun Zhang* and *Min Zhang.* Shanghai University of Finance and Economics

12:10 Floor Discussion.

**Session 43: New Methods for Survival Analysis and Network Analysis with Application to Biomarker Studies (Invited)**
Room: 302B
Organizer: Yuanjia Wang, Columbia University.
Chair: Yuanjia Wang, Columbia University.

10:30 Learning causal networks via additive faithfulness
*Kuang-Yao Lee,* *Tianqi Liu,* Bing Li and *Hongyu Zhao.* Temple University

10:55 Identifying Disease-Associated Biomarker Network Features Through Conditional Graphical Model
*Shanghong Xie,* Donglin Zeng and *Yuanjia Wang.* Columbia University

11:20 Generalized Mean Residual Life Models for Case-Cohort and Nested Case-Control Studies
*Peng Jin,* Anne Zeleniuch-Jacquotte and *Mengling Liu.* New York University

11:45 Adjusting for Handling Effects in Microarray Data for Survival Risk Prediction
*Andy Ni,* Mengling Liu and *Li-Xuan Qin.* Ohio State University

12:10 Floor Discussion.

**Session 44: Statistical Challenges at the Intersection of Prediction and Causality (Invited)**
Room: 302C
Organizer: Jordan Rodu, University of Virginia.
Chair: Jared Murray, University of Texas at Austin.

10:30 Causal inference in algorithmic fairness
*Joshua Loftus,* Matt Kasner, Ricardo Silva and Chris Russell.
New York University

11:45 Adjusting for Handling Effects in Microarray Data for Survival Risk Prediction
*Andy Ni,* Mengling Liu and *Li-Xuan Qin.* Ohio State University

12:10 Floor Discussion.
Session 45: RWE in Drug Development and Observational Studies: Examples and Discussion (Invited)
Room: 305A
Organizer: Xuanyao He, Eli Lilly and Company.
Chair: Lingsong Zhang, Purdue University.

10:30 Outcome weighted learning for imbalanced high dimensional precision medicine problems
*Hui Sun¹, Bruce Craig and Lingsong Zhang. ¹Purdue University

10:55 Multivariate spectral downscaling for speciated PM2.5
*Mikael Kusela¹, Donata Giglio², Anirban Mondal³ and Michael Stein⁴. ¹Carnegie Mellon University ²University of Colorado Boulder ³Case Western Reserve University ⁴University of Chicago

11:45 Structured Latent Factor Analysis for Large-scale Data
*Xiaou Li⁵, Yanxiao Chen² and Siliang Zhang³. ¹University of Minnesota ²London School of Economics and Politics ³Fudan University

Session 46: New Advances in High-dimensional Multivariate Analysis (Invited)
Room: 305B
Organizer: Xiaou Li, University of Minnesota.
Chair: Xiaou Li, University of Minnesota.

10:30 Estimation of a functional of high-dimensional mean and covariance matrix
*Haolei Weng¹, Yifeng Zhou² and Jianqing Fan². ¹Michigan State University ²Princeton University

11:45 Target population statistical inference with data integration across multiple sources
*Yang Song¹ and Xihao Li². ¹Vertex Pharmaceuticals, Inc. ²Harvard University

Session 47: SAMSI session on advances in spatio temporal modeling for health and environment (Invited)
Room: 306A
Organizer: Yawen Guan, North Carolina State University.
Chair: Yawen Guan, North Carolina State University.

10:30 On the Probability Distributions of Duration of Heatwaves
*Sohini Raha and Sajit Ghosh. North Carolina State University

11:20 Survival Extrapolation in Cost-effectiveness Analysis: Combining Clinical Trial and Real-World Data
*Binbing Yu, Jinfeng Xu¹ and Na Zhao. ¹AstraZeneca

11:45 Discussant: Martin Ho, U.S. Food and Drug Administration

Session 48: Real world evidence and big data (Invited)
Room: 306B
Organizer: Meijing Wu, Abbvie, Inc..
Chair: Bo Fu, Astellas Pharma.

10:30 Globally scalable quantile regression with SGD for massive real-world data
*Yixin Fang¹, Jinfeng Xu² and Na Zhao². ¹Abbvie, Inc. ²University of Hong Kong

11:20 Survival Extrapolation in Cost-effectiveness Analysis: Combining Clinical Trial and Real-World Data
*Binbing Yu. AstraZeneca

11:45 Discussant: Martin Ho, U.S. Food and Drug Administration

Session 49: Statistical Learning Advancement for Inference in Big Data Age (Invited)
Room: 306C
Organizer: Honglang Wang, Indiana University-Purdue University Indianapolis.
Chair: Honglang Wang, Indiana University-Purdue University Indianapolis.
10:30 Statistical Inference on Partially Linear Panel Model under Unobserved Linearity

*Ruqi Liu, Ben Boukai and Zhaofeng Shang.* Indiana University-Purdue University Indianapolis

10:55 Asymptotic Properties of Neural Network Sieve Estimators

*Xiaoxi Shen, Chang Jiang, Lyudmila Sakhnenko and Qing Lu.* Michigan State University

11:20 Auto-encoding graph-valued data with applications to brain connectomes

*Meimei Liu, Zhengwu Zhang and David Dunson.* 1. Duke University 2. University of Rochester

11:45 An Fast Algorithm For The A-optimal Sampling Distributions In A Big Data Linear Regression

*Hanxiang Peng, Fei Tan, Sheng Zhang and Haixia Li.* Math Sciences, IUPUI

12:10 Floor Discussion.

**Session 50: Moving towards Multimodal and Multi-layer imaging analysis (Invited)**

Room: 303
Organizer: Haochang Shou, University of Pennsylvania.
Chair: Luo Xiao, North Carolina State University.

10:30 Joint and Individual Non-Gaussian Component Analysis for Data Integration

*Benjamin Risk* and *Irina Gaynanova.* 1. Emory University 2. Texas A&M University

10:55 Methods for Analyzing Multi-modal Imaging Data


11:20 Characterizing the Longitudinal Behavior of Multiple Sclerosis Lesions on Structural Magnetic Resona


11:45 On testing for spatial correspondence between maps of human brain structure and function

*Simon Vandekar.* Vanderbilt University

12:10 Floor Discussion.

**Session 51: Recent Development of Artificial Intelligence in Healthcare (Invited)**

Room: 304
Organizer: Yajj Xu, Center for Devices and Radiological Health, FDA.
Chair: Jincuo Wu, Center for Devices and Radiological Health, FDA.

10:30 Bayesian scalar-on-image neural network with application to skin cancer images

*Jian Kang.* University of Michigan

10:55 AI/ML for Pharma R&D: Analytical Challenges and Opportunities

*Roy Liu.* Takeda Pharmaceuticals

11:20 Statistical Evaluation of Medical Devices that Incorporate Artificial Intelligence

*Yajj Xu, Xiaolin Xiong and Changhong Song.* Center for Devices and Radiological Health, FDA

11:45 Responder Analysis by Deep Learning with Application to a Phase III Clinical Trial


12:10 Floor Discussion.

**Session 52: Contributed Session 1 (Contributed)**

Room: 307
Chair: Elisavet Sofikitou, University at Buffalo.

10:30 Checking Adequacy of Variance Function with Unknown Mean Function

*Jia Liang and Weixing Song.* Kansas State University

10:50 A new gene-based association test

*Zhongxue Chen* and *Kai Wang.* 1. Indiana University 2. University of Iowa

11:10 Estimation of ROC Curve and AUC via Heterogeneous Box-Cox Transformations


11:30 Optimal two-stage design for estimating the Area under the receiver operating characteristic curve

*Yougui Wu.* University of South Florida

11:50 A Class of Bivariate Semiparametric Control Charts with Application to Biostatistics

*Elisavet Sofikitou.* University at Buffalo

12:10 Floor Discussion.

**Session 53: Statistics in Biosciences Invited Session: Statistical Advancement in Bioscience Research (Invited)**

Room: 301A
Organizer: Mei-Cheng Wang, Johns Hopkins University.
Chair: Mei-Cheng Wang, Johns Hopkins University.

14:00 New methods for estimating follow-up rates in cohort studies

*Xiaonan Xue.* Albert Einstein College of Medicine

14:25 Spatiotemporal Statistical Methods for Monitoring Glaucoma Progression Using Visual Field Data


14:50 Applying CHW Method to 2-in-1 Design: Gain or Lose?

*Xiaofei Bai* and *Qi Qi Deng.* 1. Boehringer Ingelheim, Inc

15:15 Floor Discussion.
Session 54: Innovative statistical analysis and design in handling unconventional clinical data (Invited)
Room: 301B
Organizer: Yaohua Zhang, Vertex Pharmaceuticals, Inc.
Chair: Yaohua Zhang, Vertex Pharmaceuticals, Inc.
14:00 Causal Effects in Clinical Endpoint Bioequivalence Studies in the Presence of Intercurrent Events
Yiyue Lou1, Michael Jones2 and Wanjie Sun3.
1Vertex Pharmaceuticals, Inc. 2University of Iowa 3Center for Drug Evaluation and Research, FDA
14:25 When convention meets practicality: Pooled analysis testing under the two-study paradigm
Dong Xi, Frank Bretz and Willi Maurer. Novartis
14:50 Using Surrogate Endpoints in Adaptive Design with Delayed Treatment Effect
Qing Li and Jianchang Lin. Takeda Pharmaceuticals
15:15 Seamless Phase 2/3 Oncology Trial Design with Flexible Sample Size Determination
Zhaoyang Teng1, Yuan Tian2, Yi Liu3 and Guohui Liu4.
1Takeda Pharmaceuticals 2North Carolina State University 3Nektar Therapeutics 4Takeda Pharmaceuticals
15:40 Floor Discussion.

Session 55: Recent Advances in Statistical Normalization Methods for Sequencing Data (Invited)
Room: 302A
Organizer: Li-Xuan Qin, Memorial Sloan Kettering Cancer Center.
Chair: Ai Ni, Memorial Sloan Kettering Cancer Center.
14:00 Copy number analysis of circulating cell-free DNA
Venkatraman Seshan, Nicholas Socci, Dana Tsui and Julie Yang.
Memorial Sloan Kettering Cancer Center
14:20 Super-delta: a robust method that combines data normalization and differential expression analysis
Yuhang Liu1, Jinfeng Zhang3 and Xing Qiu2.
1Florida State University 2University of Rochester
14:40 Statistical assessment of depth normalization methods for microRNA sequencing
Li-Xuan Qin. Memorial Sloan Kettering Cancer Center
15:00 SCnorm: A quantile-regression based approach for normalization of single-cell RNA-seq data
Rhonda Bacher. University of Florida
15:20 Normalization approaches for gene expression studies completed with high-throughput sequencing
Brooke Fridley1, Farnoosh Aghababazadeh1 and Qian Li2.
1Moffitt Cancer Center 2University of South Florida
15:40 Floor Discussion.

Session 56: Biomarkers in Cancer Clinical Trials (Invited)
Room: 302B
Organizer: Mei-Yin Polley, Mayo Clinic.
Chair: Ying Lu, Stanford University.
14:00 Bayesian adaptive enrichment design for continuous biomarkers
Lindsay Renfro1 and Yusha Liu2.
1University of Southern California 2University of Iowa
14:25 On the Design of Stratified Biomarker Trials With Measurement Error
Susan Halabi1, Chen-Yen Lin2 and Aiyi Liu3.
1Duke University 2Eli Lilly and Company 3National Institutes of Health
14:50 Two Stage Adaptive Design for Prognostic Biomarker Signature with Survival Outcome
Biyue Dai1 and Mei-Yin Polley2.
1University of Iowa 2Mayo Clinic
15:15 Utilizing Integrated analysis of RNA and DNA from phase III oncology trials to assess response
Terry Hyslop1, Chuck Perou2, Katherine Hoadley2 and Maki Tanioka2.
1Duke University 2University of North Carolina at Chapel Hill
15:40 Floor Discussion.

Session 57: Recent Advances in Precision Medicine Research (Invited)
Room: 302C
Organizer: Yingqi Zhao, Fred Hutchinson Cancer Research Center.
Chair: Yingqi Zhao, Fred Hutchinson Cancer Research Center.
14:00 Constructing personalized decision algorithm for mHealth applications
Xinyu Hu, Min Qian, Bin Cheng and Ken Cheung.
Columbia University
14:25 Comparative Intervention Scoring for Heterogeneity of Long-term Health System Intervention Effects
Jared Huling1, Menggang Yu2 and Maureen Smith1.
1Ohio State University 2University of Wisconsin-Madison
14:50 Optimal dynamic treatment regimes using decision lists
Yichi Zhang1, Eric Laber2, Butch Tsiatis2 and Marie Davidian2.
1University of Rhode Island 2North Carolina State University
15:15 Improved doubly robust estimation in learning individualized treatment rules
Yinghao Pan1 and Yingqi Zhao2.
1University of North Carolina at Charlotte 2Fred Hutchinson Cancer Research Center
15:40 Floor Discussion.

Session 58: Statistical Machine Learning in Analyzing Large-scale Multi-Dimensional Data (Invited)
Room: 305A
Organizer: Xiwei Tang, University of Virginia.
Chair: Xiwei Tang, University of Virginia.
14:00 A Temporal Latent Factor Model for Product Sales Forecasting
Xuan Bi1, Gediminas Adomavicius1, William Li2 and Annie Qu3.
1University of Minnesota 2Shanghai Jiao Tong University 3University of Illinois at Urbana-Champaign
Session 60: Gaussian graphical models and sparsity (Invited)

Room: 306A
Organizer: Yichao Wu, University of Illinois at Chicago.
Chair: Yichao Wu, University of Illinois at Chicago.

14:00 Sparse graphical modeling of longitudinal data
Yafei Zhang and *Pang Du. Virginia Polytechnic Institute and State University

14:25 Simultaneous Change Point Detection and Structure Recovery for Gaussian Graphical Models
Yufeng Liu. University of North Carolina at Chapel Hill

14:50 High-dimensional Q-learning for dynamic treatment regimes
Rui Song. North Carolina State University

15:15 An approximate EM algorithm for latent Gaussian graphical model
*Chaowen Zheng and Yichao Wu. 1North Carolina State University 2University of Illinois at Chicago

15:40 Floor Discussion.

Session 61: Advanced methods in high dimensional and big data analysis (Invited)

Room: 306B
Organizer: Yichuan Zhao, Georgia State University.
Chair: Pengsheng Ji, University of Georgia.

14:00 On post dimension reduction statistical inference
*Bing Li, Kyongwon Kim, Lexin Li and Zhou Yu. 1Pennsylvania State University

14:25 Copula-based semiparametric analysis for time series data with detection limits
Fuyuan Li, Yanlin Tang and *Huixia Judy Wang. 1East China Normal University 2George Washington University

14:50 Big EHR Data Analysis and Modeling: Challenges and Opportunities
Hulin Wu. University of Texas Health Science Center at Houston

15:15 Floor Discussion.

Session 62: Recent Advances in Statistical Methodology for Analyzing Human Metgenomics Data (Invited)

Room: 306C
Organizer: Ni Zhao, Johns Hopkins University.
Chair: Ni Zhao, Johns Hopkins University.

14:00 Powerful adaptive microbiome differential abundance analysis
Xiang Zhan. Pennsylvania State University College of Medicine

14:25 HMPSData: Integrative Human Microbiome Data R Bioconductor package and analysis workflow
Ekaterina Smirnova, Ni Zhao, Mikhail Dozmorov and Levi Waldron. 1Virginia Commonwealth University 2Johns Hopkins University 3CUNY Graduate School of Public Health and Health Policy

14:50 A framework for multi-view analysis of microbiome data
Timothy Randolph, Yue Wang, Ali Shojaie and Jing Ma. 1Fred Hutchinson Cancer Research Center 2University of Washington

15:15 A robust distance-based kernel association test for correlated microbiome data
Hyeunwook Koh, Yutong Li, Xiang Zhan, Jun Chen and *Ni Zhao. 1Johns Hopkins University 2University of Pennsylvania 3Pennsylvania State University 4Mayo Clinic

15:40 Floor Discussion.

Session 63: How can statisticians help pediatric drug/medical device development using innovative statistical methods? (Invited)

Room: 303
Organizer: Mary Zhao, Boehringer Ingelheim, Inc.
Chair: Mary Zhao, Boehringer Ingelheim, Inc.
Session 65: Making sense of data for business: the role of statistics (Invited)

Room: 307
Organizer: Dungang Liu, University of Cincinnati.
Chair: Dungang Liu, University of Cincinnati.

14:00 Factor models for matrix-valued high-dimensional time series
Yiwen Liu. San Diego State University

14:25 Integrative interaction analysis using threshold gradient directed regularization
Yan Li, Long Liu, Yichen Qin, and Hongjian Shi. University of North Carolina at Charlotte

14:50 Post-selecting inference for regression models with linear constraints
Yichen Qin. University of North Carolina at Charlotte


15:40 Floor Discussion.

Session 66: Meet Both the Computational and Statistical Challenges in Learning (Invited)

Room: 301A
Organizer: Wen Zhou, Colorado State University.
Chair: Wen Zhou, Colorado State University.

16:00 Parameterized matrix factorization with missing data via nonconvex optimization
Xiaodong Li. University of California, Davis

16:25 High-dimensional Gaussian graphical model on network-linked data
Tianxi Li, Cheng Qian, Elizaveta Levina, and Ji Zhu. University of Michigan

16:50 Magic Cross-Validation for Support Vector Machines and Related Large Margin Classifiers
Boxiang Wang and Hui Zou. University of Iowa

17:15 High dimensional independence testing with maxima of rank correlations
Mathias Drton, Fang Han, and Hongjian Shi. University of North Carolina at Chapel Hill

17:40 Floor Discussion.

Session 67: Identification of biomarkers and disease mechanisms through proper biostatistics analysis (Invited)

Room: 301B
Organizer: Yushi Liu, Eli Lilly and Company.
Chair: Gang Han, Texas A&M University.

16:00 Characterizing the Assay Reproducibility of a Companion Diagnostics Test
Alan Chiang. Celgene

16:25 Analysis of Flow Cytometry Data
Shuguang Huang. Stat4ward, LLC

16:50 Post-selection Inference for Regression Models with Linear Constraints
Jiarui Lu and Hongzhe Li. University of North Carolina at Chapel Hill

17:15 Bayesian Inference of Multivariate Mixed-effects Joint Models for Longitudinal-survival Data
Xan Xu and Yangxin Huang. University of South Florida

17:40 Floor Discussion.
Session 68: Statistical Challenges and Advances in Biomarker Studies (Invited)
Room: 302A
Organizer: Yuanjia Wang, Columbia University.
Chair: Fei Gao, University of Washington.

16:00 Statistical Methods for Genetic Risk Prediction
Yiming Hu, Qiongshi Lu, Mo Li, Yixuan Ye, Wei Jiang and Hongyu Zhao. Yale University 2University of Wisconsin Madison

16:25 New Directions in image-based analysis of biomarkers
Debashis Ghosh, Xuhong Zhang, Fayong Xing and Olivier Simon. Colorado School of Public Health

16:50 Using Information Theory to Evaluate Surrogacy of Biomarkers
Maria Carmen, Hua Jin, Qian Zhao and Ying Lu. Complutense University of Madrid 2South China Normal University 3Guangzhou Medical University 4Stanford University

17:15 Identifying Biomarker Networks From Multiple Modality Data
Yanjia Wang, Shanghong Xie and Donglin Zeng. Columbia University 2University of North Carolina at Chapel Hill

17:40 Floor Discussion.

Session 69: Modeling and inference for complex dependence in large datasets (Invited)
Room: 302B
Organizer: Kai Zhang, University of North Carolina at Chapel Hill.
Chair: Ding-Geng Chen, University of North Carolina at Chapel Hill.

16:00 How to Measure the Similarity of Data?
Gabor Szekely. National Science Foundation

16:25 Multi-scale testing of independence and conditional independence for massive data
Shai Gorsky, Jialiang Mao and Li Ma. Duke University

16:50 Detect with BERET
Duveol Lee, Kai Zhang and Michael Kosorok. University of North Carolina at Chapel Hill

17:15 Essential Regression
Xin Bing and Florentina Bunea. Cornell University

17:40 Floor Discussion.

Session 70: Contemporary Viewpoints on Improving Adaptive Phase I Clinical Trial Designs (Invited)
Room: 302C
Organizer: Tom Braun, University of Michigan.
Chair: Tom Braun, University of Michigan.

16:00 How Many Patients Should We Enroll in an Adaptive Phase I Trial?
Thomas Braun. University of Michigan

16:25 Phase I designs that allow for uncertainty in the attribution of adverse events
Alexia Iasonos and John Oquigley. Memorial Sloan Kettering Cancer Center 2University Paris 6

16:50 Shift models for dose finding in partially ordered groups
Bethany Horton, Nolan Wages and Mark Conaway. University of Virginia

17:15 Combining Data from a Phase I-II dose Escalation Study and a Follow-up Cohort
Joseph Koopmeiners, Kasha Mohammad and Jeffrey Boatman. University of Minnesota 2University of Texas

17:40 Floor Discussion.

Session 71: Advances in High-Dimensional Inference and Regression (Invited)
Room: 305A
Organizer: Jessie Jeng, North Carolina State University.
Chair: Jessie Jeng, North Carolina State University.

16:00 An Empirical Bayes Approach for Selection Bias in Functional Data
Joshua Derenski, Yingying Fan and Gareth James. University of Southern California

16:25 RANK: Large-Scale Inference with Graphical Nonlinear Knockoffs
Yingying Fan, Emre Demirkaya, Gaorong Li and Jinchu Lv. University of Southern California 2Beijing University of Technology

16:50 Effective Joint Modeling through the Multivariate Debiased Lasso
Jacob Rynne. North Carolina State University

17:15 Discussant: Jessie Jeng. North Carolina State University

17:40 Floor Discussion.

Session 72: Subgroup identification techniques and practical applications in drug development (Invited)
Room: 305B
Organizer: Lei Xu, Vertex Pharmaceuticals, Inc..
Chair: Lei Xu, Vertex Pharmaceuticals, Inc..

16:00 Overview of methods for subgroup identification in clinical trials
Ilya Lipkovich and Alex Dmitrienko. Eli Lilly and Company 2Mediana, Inc.

16:25 Subgroup Identification Through Confident Inference on Treatment Efficacy for Time-to-Event Outcomes
Yue Wei and Ying Ding. University of Pittsburgh

16:50 Subgroup Identification and Multi-population Tailoring Studies: Moving from Phase II to Phase III
Ying Zhang and Lei Shen. Eli Lilly and Company

17:15 Floor Discussion.

Session 73: Recent Advances in joint modeling and dynamic prediction on longitudinal and survival data (Invited)
Room: 306A
Organizer: Cong Xu, Vertex Pharmaceuticals, Inc..
Chair: Cong Xu, Vertex Pharmaceuticals, Inc..
16:00 Assessing predictive accuracy of survival regressions subject to non-independent censoring

Ming Wang, Qi Long, Chixiang Chen and Lijun Zhang. Pennsylvania State University and University of Pennsylvania.

16:25 A class of models for dynamic prediction of survival using multivariate longitudinal data

Liang Li. University of Texas MD Anderson Cancer Center.

16:50 Landmark analysis for the effects of longitudinal cholesterol profiles on the risk of CHDs

Bin Shi, Peng Wei and Xueling Huang. University of Texas and University of Texas MD Anderson Cancer Center.

17:15 Floor Discussion.

Session 74: Recent advances in modern nonparametric statistics with applications in ROC curve analysis (Invited)

Room: 306B
Organizer: Dongliang Wang, SUNY Upstate Medical University
Yichuan Zhao, Georgia State University.
Chair: Chi Song, Ohio State University.

16:00 Nonparametric Mass Imputation for data integration

Sixia Chen. University of Oklahoma Health Sciences Center.


Dungang Liu, Shaobo Li and Yan Yu. University of Cincinnati and University of Kansas.

16:50 Bandwidth-free ROC Curve Estimation via Bernstein Polynomial

Dongliang Wang and Xueya Cai. SUNY Upstate Medical University and University of Rochester.

17:15 Floor Discussion.

Session 75: Recent Advances in Neuroimaging Analysis (Invited)

Room: 306C
Organizer: Lexin Li, University of California, Berkeley.
Chair: Dehan Kong, University of Toronto.

16:00 Defining the resolution of optogenetic circuit mapping

Shizhe Chen, Ben Shababo, Karl Kilborn, Xinyi Deng, Johannes Friedrich, Hillel Adesnik and Lian Paninski. University of California, Davis.

16:25 Sparse Generalized Eigenvalue Problem with an Application to Neuroscience

Kean Ming Tan. University of Minnesota.

16:50 Covariate Assisted Principal Regression for Covariance Matrix Outcomes

Yi Zhao, Bingkai Wang, Stewart Mostofsky, Brian Caffo and Xi Luo. Johns Hopkins University and University of Texas.

17:15 Individualized Multilayer Tensor Learning with An Application in Imaging Analysis

Xiwei Tang, Xuan Bi and Annie Qu. University of Virginia and University of Minnesota and University of Illinois at Urbana-Champaign.

17:40 Floor Discussion.

Session 76: The Jiann-Ping Hsu Invited Session on Biostatistical and Regulatory Sciences (Invited)

Room: 303
Organizer: Lili Yu, Georgia Southern University.
Chair: Congjian Liu, Georgia Southern University.

16:00 A measure and cut-point selection criterion for k-stage diseases using Kullback-Leibler divergence.

Chen Mo, Han Samawi, Jingjing Yin and Jing Kersey. Georgia Southern University.

16:25 Misclassification Simulation Extrapolation in survival analysis with AFT model

Congjian Liu and Lili Yu. Georgia Southern University.

16:50 Regularization in Accelerated Failure Time (AFT) models with frailty parameters


17:15 A Quantitative Assessment of Risk for Subgroup Pursuit in Clinical Trials

Xinzhou Guo, Ruosha Li and Xuming He. University of Texas Health Science Center at Houston.

17:40 Floor Discussion.

Session 77: High Dimensional Statistical Modelling of Genomics (Invited)

Room: 304
Organizer: Honglang Wang, Indiana University-Purdue University Indianapolis.
Chair: Honglang Wang, Indiana University-Purdue University Indianapolis.

16:00 A Kernel-Based Neural Network for High-dimensional Genetic Risk Prediction Analysis

Xiaoxi Shen, Xiaoran Tong and Qing Lu. Michigan State University.

16:25 Genetic Heritability and Coefficients of Determination

Dabao Zhang. Purdue University.

16:50 High Dimensional Mediation Analysis with Applications to Causal Gene Identification

Qi Zhang. University of Nebraska.

17:15 Statistical modeling of genomics and epigenomics big-data reveals human disease mechanisms


17:40 Floor Discussion.

Session 78: Contributed Session 2 (Contributed)

Room: 307
Chair: Yifan Sun, Renmin University.

16:00 A Bayesian adaptive marker-stratified design with customized hierarchical modeling

Yong Zhang, Beihei Guo, Yan Han, Sha Cao and Chi Zhang. Indiana University and Louisiana State University.

17:15 Statistical modeling of genomics and epigenomics big-data reveals human disease mechanisms

June 12 08:30 - 10:10

Session 79: Recent advance in precision medicine methods
(Invited)
Room: 301A
Organizer: Min Qian, Columbia University.
Chair: Min Qian, Columbia University.

8:30 Comparing Multiple Adaptive Interventions under General SMART Designs
♦ Xiaobo Zhong1, Bin Cheng2, Min Qian2 and Ying Ken Cheung2. 1Icahn School of Medicine at Mount Sinai 2Columbia University

8:55 Constructing Stabilized Dynamic Treatment Regimes for Censored Data
♦ Yingzi Zhao1, Ruoming Zhu2, Guanhua Chen3 and Yingye Zheng1. 1Fred Hutchinson Cancer Research Center 2University of Illinois at Urbana-Champaign 3University of Wisconsin Madison

9:20 Stochastic Tree Search for Estimating Optimal Dynamic Treatment Regimes
♦ Lu Wang and Yilin Sun. University of Michigan

9:45 On Estimation of Optimal Treatment Regimes for Maximizing Quality Adjusted Survival Time
Sun Hat, ♦Ashkan Ertefaie and Brent Johnson. University of Rochester

10:10 Floor Discussion.

Session 80: Structure learning and statistical inference for complex data analysis (Invited)
Room: 301B
Organizer: Weining Shen, University of California, Irvine.
Chair: Weining Shen, University of California, Irvine.

8:30 Nonparametric conditional density estimation based on pooled biomarker assessments
Xichen Mou, ♦Dewei Wang and Joshua Tebbs. University of South Carolina

8:55 Generalized probabilistic principal component analysis of correlated data
♦ Mengyang Gu1 and Weining Shen2. 1Johns Hopkins University 2University of California, Irvine

9:20 A Fractional Imputation Procedure for Complex Surveys
Pushpal Mukhopadhyay. SAS Institute, Inc.

9:45 Logistic Regression Augmented Community Detection for Network Data
♦ Yunpeng Zhao1, Qing Pan2 and Chengan Du3. 1Arizona State University 2George Washington University 3Yale University

10:10 Floor Discussion.

Session 81: Recent Advances of Statistical Modeling in Biomedical Research (Invited)
Room: 302A
Organizer: Dehan Kong, University of Toronto.
Chair: Yifan Cui, University of Pennsylvania.

8:30 Optimal designs of two-phase studies
Ran Tao. Vanderbilt University Medical Center

8:55 An Analysis of Classical Multidimensional Scaling
♦ Yuying Xie1, Anna Little1 and Qiang Sun2. 1Michigan State University 2University of Toronto

9:20 Learning Optimal Individualized Decision Rules with Risk Control
♦ Zhengling Qi1, Jong-Shi Pang2 and Yufeng Liu3. 1University of North Carolina at Chapel Hill 2University of Southern California

9:45 Floor Discussion.

Session 82: Recent Methods for Complex Survival Data (Invited)
Room: 302B
Organizer: Jiajia Zhang, University of South Carolina.
Chair: ALEXANDER MCLAIN, University of South Carolina.

8:30 Quantile Association Regression on Bivariate Survival Data
♦ Ling-Wan Chen1, Yu Cheng2, Ying Ding2 and Ruoshia Li3. 1National Institute of Environmental Health 2University of Pittsburgh 3University of Texas Health Science Center at Houston

8:55 A Varying-Coefficient Generalized Odds Rate Model with Time-Varying Exposure
Jie Zhou1, ♦Jiayia Zhang2, Alexander Mclain1, Wenbin Lu3, Xuemei Su4 and James Hardin1. 1University of South Carolina 2University of South Carolina 3North Carolina State University

9:20 Gene-based Association Test for Bivariate Survival Data via Functional Regression with Copula Models
Ying Ding. Department of Biostatistics, University of Pittsburgh

9:45 Joint analysis of interval-censored recurrent events and survival data
Guangli Yu1, ♦Yang Li2, Liang Zhu3, Hui Zhao1, Jia-guo Sun2 and Leslie Robison2. 1Eli Lilly and Company

2019 ICSA Applied Statistics Symposium, Raleigh, NC, June 9-12

56
Session 83: New statistical methods for big data in biomedical research (Invited)
Room: 302C
Organizer: Quefeng Li, University of North Carolina at Chapel Hill.
Chair: Quefeng Li, University of North Carolina at Chapel Hill.
8:30 Combinatorial Inference
*Junwei Lu*, *Matey Neykov* and *Han Liu*. 1Harvard University 2Carnegie Mellon University 3Northwestern University
8:55 CORALS: Co-clustering analysis via regularized alternating least squares
*Gen Li*. Columbia University
9:20 CESME: Cluster Analysis with Latent Semiparametric Mixture Models
*Wen Zhou*, *Lyuou Zhang*, *Lulu Wang* and *Hui Zou*. 1Colorado State University 2Gilead Sciences, Inc. 3University of Minnesota
9:45 Robust high-dimensional asymmetric data analysis with non-convex mean regression
*Bin Luo and Xiaoli Gao*. University of North Carolina at Greensboro
10:10 Floor Discussion.

Session 84: Case Studies and Methods for Learning and Improving Healthcare Through EHRs (Invited)
Room: 305A
Organizer: Qingxia Chen, Vanderbilt University Medical Center.
Chair: Yu Cheng, University Of Pittsburgh.
8:30 Repurposing Statistical Tools to Improve Data Quality for Time-varying Measurements in EHRs
*Qingxia Chen and Trent Rosenbloom*. Vanderbilt University Medical Center
8:55 PIE: A prior knowledge guided estimation method for bias reduction in EHR-based research
*Jing Huang*, *Rui Duan*, *Rebecca Hubbard*, *Yonghui Wu*, *Jason Moore*, *Hua Xu* and *Yong Chen*. 1University of Pennsylvania 2University of Florida 3University of Texas at Houston Health Science Center
9:20 Missing at random or not: a semiparametric testing approach via instrumental variable
*Rui Duan*, *Jason Liang*, *Pamela Shaw*, *Chengyong Tang* and *Yong Chen*. 1University of Pennsylvania 2National Institute of Allergy and Infectious Disease 3Temple University
9:45 Recalibrating Prognostic Risk Score Adapted to EHR Data
*Dandan Liu, Qingxia Chen, Hui Nian and Joshua Denny*. Vanderbilt University
10:10 Floor Discussion.

Session 85: Longitudinal, Multilevel, Multiway and Spatial Functional Data Analysis (Invited)
Room: 305B
Organizer: Luo Xiao, North Carolina State University.
Chair: Luo Xiao, North Carolina State University.
8:30 Modeling continuous glucose monitoring (CGM) data during sleep
*Ciprian Crainiceanu*. Johns Hopkins University
8:55 Dynamic Prediction of Alzheimer's Disease Using Multiple Longitudinal Outcomes and Survival Data
*Kan Li* and *Sheng Luo*. 1Merck Research Laboratory, Merck & Co. 2Duke University
9:20 Tests of significance for time-varying covariate effect in longitudinal functional data
*Ana-Maria Staicu*, *Saeblina Oh* and *Arnab Maity*. 1Carolina State University 2North Carolina State University
9:45 Covariance Function Estimation for Multidimensional Functional Data
*Jiayi Wang*, *Raymond K. W. Wong* and *Xiaoke Zhang*. 1Texas A& M University 2George Washington University
10:10 Floor Discussion.

Session 86: Random forests and decision trees for survival data (Invited)
Room: 306A
Organizer: Ruoning Zhu, University of Illinois at Urbana-Champaign.
Chair: Ruoning Zhu, University of Illinois at Urbana-Champaign.
8:30 Subgroup Identification using Covariate Adjusted Interaction Trees
*Jon Steingrimsson and Jiabei Yang*. Brown University
8:55 Oblique Random Survival Forest
*Byron Jaeger*, *Leann Long*, *Dustin Long*, *Leslie Mcclure*, *Jeff Szczukowski*, *George Howard*, *Noah Simon*, *Yuan-I Min* and *Mario Sims*. 1University of Alabama at Birmingham 2Drexel University 3University of Mississippi Medical Center
9:20 Treatment decision using causal survival forests
*Yifan Cui*, *Michael Kosorok*, *Stefan Wager* and *Ruoping Zhu*. 1University of Pennsylvania 2University of North Carolina at Chapel Hill 3Stanford University 4University of Illinois at Urbana-Champaign
9:45 Penalized Random Survival Forests
*Sarah Formentini and Ruoping Zhu*. University of Illinois at Urbana-Champaign
10:10 Floor Discussion.

Session 87: New developments in nonparametric statistics and empirical likelihood (Invited)
Room: 306B
Organizer: Yichuan Zhao, Georgia State University.
Chair: Wei Zheng, University of Tennessee at Knoxville.
Session 88: Statistical Controversies in Forensic Evidence Interpretation (Invited)
Room: 306C
Organizer: Jan Hannig, University of North Carolina at Chapel Hill.
Chair: Jonathan Williams, University of North Carolina at Chapel Hill.

8:30 The Critical Role of Statistics in Evaluating Forensic Evidence
Karen Kafadar. University of Virginia

8:55 Statistical Approaches for Analyzing and Presenting Forensic Footwear Impression Evidence
Steven Lund. National Institute of Standards and Technology

9:20 Which statistical paradigm should I use for forensic evidence interpretation?
Danica Ommen1 and Christopher Saunders2. 1Iowa State University 2South Dakota State University

9:45 Are reported likelihood ratios well calibrated?
Jan Hannig2 and Hari Iyer2. 1University of North Carolina at Chapel Hill 2National Institute of Standards and Technology

10:10 Floor Discussion.

Session 89: Latent attribute models and their applications (Invited)
Room: 303
Organizer: Gongjun Xu, University of Michigan.
Chair: Gongjun Xu, University of Michigan.

8:30 Structured Latent Class Regression Analysis of Multivariate Binary Data: Estimating Disease Etiology
Zhenke Wu. University of Michigan

8:55 Estimation of Q-matrix with unknown number of attributes
Yinghan Chen1, Ying Liu2, Steven Culpepper2 and Yuguoke Chen2. 1University of Nevada-Reno 2University of Illinois at Urbana-Champaign

9:20 Learning Attribute Patterns in High-Dimensional Structured Latent Attribute Models
Yuqi Gu and Gongjun Xu. University of Michigan

9:45 Identified Bayesian Factor Analysis
Albert Man and Steven Culpepper. University of Illinois at Urbana-Champaign

10:10 Floor Discussion.

Session 90: Recent Development in High-Dimensional Data Analysis (Invited)
Room: 304
Organizer: Xiaoli Kong, Loyola University Chicago.
Chair: Xiaoli Kong, Loyola University Chicago.

8:30 A two sample test of equality of means in high dimensional data
Huiyu Zhang and Haiyan Wang. Kansas State University

8:55 High-Dimensional Rank-Based Inference
Solomon Harrar1 and Xiaoli Kong2. 1University of Kentucky 2Loyola University Chicago

9:20 Sufficient dimension reduction using a divergence measure for high dimensional data
Qingcong Yuan1, Wenhui Sheng2 and Xiangrong Yin. 1University of Virginia 2Marquette University 3University of Kentucky

9:45 On sparse Fourier transform inverse regression for sufficient variable selection
Jianying Weng and Xiangrong Yin. University of Kentucky

10:10 Floor Discussion.

Session 91: Novel genetic discovery and risk prediction methods for Alzheimer disease (Invited)
Room: 307
Organizer: Yi-Ju Li, Duke University Medical Center.
Chair: Yi-Ju Li, Duke University Medical Center.

8:30 Family-based association tests for rare variants with censored traits
Wenjing Qi1, Shumian Sun1, X. Raymond Gao2, Eden R. Martin3, Andrew S. Allen1 and Yi-Ju Li3. 1Duke University 2Ohio State University 3University of Miami

8:55 Precise modelling zero-inflated count phenotypes from sequencing data
Qiao Fan1, Shumian Sun2 and Yi Ju Li2. 1Centre for Quantitative Medicine, Duke-NUS 2Duke University Medical Center

9:20 Post-GWAS data integration identifies risk factors for Alzheimer’s disease
Qiongshi Lu. University of Wisconsin-Madison

9:45 Evaluation of polygenic risk scores for prediction of complex disease
X. Raymond Gao1, Yi-Ju Li2 and Eden Martin3. 1Ohio State University 2Duke University 3University of Miami

10:10 Floor Discussion.
June 12 10:30 - 12:10

Session 92: Bayesian methodology and applications in clinical trials (Invited)
Room: 301A
10:30 A simple approach to incorporating historical data in clinical trial design and analysis
• Lanju Zhang, Zailong Wang, Li Wang, Lu Cui, Jeremy Sokolove and Ivan Chan, Abbvie, Inc.
10:55 Bayesian logistic mixed effects model for multiple imputation of frequently repeated binary outcomes
• Tianmeng Lyu, Björn Holzhauer, Dong Xi and Alkaz Uddin. Novartis Pharmaceuticals Corporation
11:20 Detecting Differentially Methylated Regions Accounting for Cell Purity Using Bayesian Credible Band Shili Lin. Ohio State University
11:45 Compartment models using PROC NLMIXED Raghavendra Rao Kurada. SAS Institute, Inc.
12:10 Floor Discussion.

Session 93: Statistical Methods for Integrative Analysis of Multi-Omics Data (Invited)
Room: 301B
Organizer: Shaoyu Li, University of North Carolina at Charlotte, Chair: Shaoyu Li, University of North Carolina at Charlotte.
10:30 Robust integrative data analysis using penalized weighting methods Xiaoli Gao. University of North Carolina at Greensboro
10:55 An efficient prospective likelihood approach to secondary traits-genetic association analyses
• Guolian Kang1, Wenzhen Bi3, Yan Li2, Matthew Smeltzer3, Guimin Gao4 and Shengli Zhao5. 1St. Jude Children’s Research Hospital 2University of North Carolina at Chapel Hill 3University of Memphis 4University of Chicago 5Qufu Normal University
11:20 Robust integration of multi-omics data for Gene-Environment Interactions Xi Lu and Cen Wu1. 1Kansas State University
11:45 Floor Discussion.

Session 94: Innovative methods for handling missing data in the era of data science (Invited)
Room: 302A
Organizer: Jiwei Zhao, State University of New York at Buffalo. Chair: Jiwei Zhao, State University of New York at Buffalo.
10:30 Bayesian profiling multiple imputation for missing electronic health records
• Yajuan Si1, Mari Palta2 and Maureen Smith2. 1University of Michigan 2University of Wisconsin-Madison
10:55 A simple Index of Local Sensitivity to Non-Ignorability for Intensive Longitudinal Data with missing
• Hui Xie1, Chengbo Yuan2, Donald Hedeker3 and Robin Mermelstein2. 1Simon Fraser University 2University of Illinois at Chicago 3University of Chicago
11:20 Monte Carlo Approach to Likelihood Maximization for Missing Data Problems Hua Yan Chen. University of Illinois at Chicago
11:45 Robust Estimation of the Average Treatment Effect under Data Combination BaoLuo Sun. National University of Singapore
12:10 Floor Discussion.

Session 95: Statistical learning of modern data objects (Invited)
Room: 302B
Organizer: Gen Li, Columbia University.
Chair: Gen Li, Columbia University.
10:30 Accounting for cell type heterogeneity in proteogenomic association analyses
• Xiaoyu Song1, Jiayi Ji1, Lin Chen2 and Pei Wang3. 1Icahn School of Medicine at Mount Sinai 2University of Chicago 3McGill University
10:55 A Unified Approach to Sparse Tweedie Modeling of Multi-Source Insurance Claim Data
• Yuwen Gu1, Simon Fontaine2, Yi Yang3, Wei Qian3 and Bo Fan2. 1University of Connecticut 2University of Montreal 3University of Delaware
11:20 Identifiability of Restricted Latent Class Models Gongjun Xu and Yuqi Gu. University of Michigan
11:45 Floor Discussion.

Session 96: Recent Advances in Survival Methods for Clinical Trials (Invited)
Room: 302C
Organizer: Li Chen, University of Kentucky.
Chair: Li Chen, University of Kentucky.
10:30 Biomarker-integrated Clinical Trials with Threshold Selection and Enrichment
• Xiaofei Wang1, Ting Wang2 and Stephen George1. 1Duke University Medical Center 2University of North Carolina at Chapel Hill
10:55 Time-to-event continual reassessment method (TITE-CRM) for bivariate outcomes of toxicity/efficacy Donglin Yan3, Nolan Wages2, Christopher Tai3, Tamila Kindwall-Keller2 and Emily Dressler4. 2University of Virginia 3PRA Health Sciences 4Wake Forest University
11:20 Optimal Two-Stage Phase II Survival Trial Design Jianrong Wu, Li Chen, Jing Wei and Heidi Weiss. University of Kentucky
11:45 Phase II trial design with growth modulation index as the primary endpoint Jianrong Wu, Li Chen, Jing Wei, Heidi Weiss, Rachel Miller and John Villano. University of Kentucky
12:10 Floor Discussion.
Session 97: New developments in big data science and biomedical applications (Invited)
Room: 305A
Organizer: Heping Zhang, Yale University
Yichuan Zhao, Georgia State University.
Chair: Guoqing Diao, George Mason University.
10:30 Detecting Dense and Sparse Signals in Biomedical Applications
*Chi Song, Lo-Bin Chang and Xiaoya Cai. Ohio State University
10:55 Multivariate Longitudinal Analysis with Functional Copula Models
Colin O. Wu. National Institutes of Health
11:20 A unified machine learning approach to estimate the clinically meaningful changes
Jiwei Zhao. State University of New York at Buffalo
11:45 Floor Discussion.

Session 98: New estimation, model selection and optimization for non-traditional data (Invited)
Room: 305B
Organizer: Annie Qu, University of Illinois at Urbana-Champaign.
Chair: Xuan Bi, University of Minnesota.
10:30 IMPUTED FACTOR REGRESSION FOR HIGH-DIMENSIONAL BLOCK-WISE MISSING DATA
*Yaqing Zhang1, Annie Qu2 and Niansheng Tang1. 1Yunnan University 2University of Illinois at Urbana-Champaign
10:55 Integrating multi-source block-wise missing data in model selection
*Fei Xue and Annie Qu. University of Illinois at Urbana-Champaign
11:20 A Parsimonious Personalized Dose Finding Model via Dimension Reduction
*Wenzhuo Zhou and Ruqing Zhu. University of Illinois at Urbana-Champaign
11:45 Floor Discussion.

Session 99: Recent Advances in Latent Variable Modeling (Invited)
Room: 306A
Organizer: XIAOJING WANG, University of Connecticut.
Chair: XIAOJING WANG, University of Connecticut.
10:30 An Exploration of Latent Structure in Process Data
10:55 A new latent variable model for the analysis of multiple partially ordered responses
Edward Ip. Wake Forest University
11:20 Skew Testlet IRT Model under a Bayesian Approach
Sandra Flores1, Caio Azevedo1, *Jorge L Bazén2 and Dipak Dey1. 1University of São Paulo 2University of Campinas
11:45 Floor Discussion.

Session 100: How Data Science Drives Success in the Enterprises (Invited)
Room: 306B
Organizer: Bo Zhang, IBM.
Chair: Bo Zhang, IBM.
10:30 Journey to an AI future
*Bo Zhang. IBM
10:55 Deploy and monitor machine learning models in production
*Liwei Wang, Ke Zhu, Matthew Neal and Emma Dickson. IBM
11:20 A progressive journey of applying data science in client advocacy
Dan Yang. IBM
11:45 build an open source data lake for data scientists
*Ke Zhu, Matthew Neal, Emma Dickson, Liwei Wang and Justin Eyster. IBM
12:10 Floor Discussion.

Session 101: New development in statistical methodologies for analyzing big neuroimaging data (Invited)
Room: 306C
Organizer: Mihye Ahn, University of Nevada-Reno.
Chair: Mihye Ahn, University of Nevada-Reno.
10:30 Nonparametric Matrix Response Regression with Application to Brain Imaging Data Analysis
Wei Hu, *Dehan Kong and Weining Shen. 1University of California, Irvine 2University of Toronto
10:55 Functional Regression Analysis of Distributional Data using Quantile Functions
*Hojin Yang1, Veerabhadran Baladandayuthapani2 and Jeffrey Morris3. 1University of Texas MD Anderson Cancer Center 2University of Michigan
11:20 Zero-Inflated Regime-Switching Stochastic Differential Equation Models for Highly Unbalanced Multiva
*Zhaohua Lu1, Sy-Min Chow2, Nilam Ram2 and Pamela M. Cole2. 1St. Jude Children’s Research Hospital 2Pennsylvania State University
11:45 Dimension deduction method for group analysis of functional neuroimaging data
Mihye Ahn. University of Nevada-Reno
12:10 Floor Discussion.

Session 102: Contributed Session 3 (Contributed)
Room: 303
Chair: Jun Young Park, University of Minnesota.
10:30 Sparse SIR: Optimal Rates and Adaptive Estimation
*Kai Tan1, Lei Shi2 and Zhou Yu3. 1University of Kentucky 2Fudan University 3East China Normal University
10:55 Robust Moderately Clipped LASSO for Simultaneous Outlier Detection and Variable Selection
*Yang Peng and Xiaoli Gao. University of North Carolina at Greensboro
Scientific Program (*Presenting Author*)

11:20 Online Updating of Information Based Model Selection Criterion in the Big Data Setting  
*Yishu Xue and Guanyu Hu*. University of Connecticut

11:45 Integrative Factorization of Bidimensionally Linked Matrices  
*Jun Young Park and Eric Lock*. University of Minnesota

12:10 Floor Discussion.

**Session 103: Modern Statistical Methods for Complex Data**  
*(Invited)*

Room: 304  
Organizer: Xinyi Li, SAMSI.  
Chair: Xinyi Li, SAMSI.

10:30 CHAMP: Post-Processing Partitions for Modularity Optimization  
*Peter Mucha*. University of North Carolina at Chapel Hill

10:55 Estimating Plant Growth Curves and Derivatives by Modeling Crowdsourced Imaged-based Data  
Haozhe Zhang\(^1\), *Dan Nettleton\(^1\), Stefan Hey\(^1\), Talukder Jubery\(^1\), Cheng-Ting Yeh\(^1\), Zihao Zheng\(^1\) and Pat Schnable\(^1\).  
\(^1\)Iowa State University

11:20 Simultaneous Confidence Corridors for Mean Functions in Functional Data Analysis of Imaging Data  
*Yueying Wang\(^1\), Guannan Wang\(^2\), Li Wang\(^3\) and R. Todd Ogden\(^3\).  
\(^1\)Iowa State University, \(^2\)College of William & Mary, \(^3\)Columbia University

11:45 An adaptive estimation in dimension reduction  
*Qin Wang*. University of Alabama

12:10 Floor Discussion.
Abstracts

Keynote Lecture I

Leveraging Real-World Data for Regulatory Decision-Making: Propensity Score-Based Patient Selection and Augmentation in Investigational Clinical Studies

Lilly Yue
U.S. Food and Drug Administration

In medical product development, there has been an increased interest in utilizing real-world data which have become abundant owing to advances in biomedical science, information technology and engineering. High-quality real-world data may be utilized to generate real-world evidence for regulatory or healthcare decision-making. This presentation will discuss two propensity score-based methods for leveraging patients from a real-world data source: one for constructing a control group for a comparative clinical study, and the other for augmenting a single-arm clinical study. The proposed approaches use the propensity score methodology to identify real-world patients who are similar to those enrolled into the investigational clinical study in terms of baseline characteristics. Either frequentist or Bayesian method can then be applied to outcome data analysis, with the option of down-weighting information from real-world data source. Examples based on pre-market regulatory review experience are provided to illustrate the implementation of the proposed approaches.

Session 1: Modern Methods for Neuroimaging Data

Tensor-on-tensor regression

Eric Lock
University of Minnesota

In neuroimaging analysis and other fields, both predictors and outcomes can take the form of a multi-way array (i.e., a tensor). We propose a framework for the linear prediction of a multi-way array from another multi-way array of arbitrary dimension, using the contracted tensor product. This framework generalizes several existing approaches, including methods to predict a scalar outcome from a tensor, a matrix from a matrix, or a tensor from a scalar. We describe an approach that exploits the multiway structure of both the predictors and the outcomes by restricting the coefficients to have reduced CP-rank. We propose a general and efficient algorithm for penalized least-squares estimation, which allows for a ridge (\(L_2\)) penalty on the coefficients. The objective is shown to give the mode of a Bayesian posterior, which motivates a Gibbs sampling algorithm for inference. We illustrate the approach with an application to metabolite resonance spectroscopy data.

A Multivariate Generalized Linear Mixed Model for Joint Modeling of Cognitive and Neuroimaging Outcome

♦ Malugeta Gebregziabher1, Carter Allen1, Daniel Baer1, Aastha Khatiwada1, Virginia Shipes1, Abbeh Teklehaimanot1 and Philip Insel1

Medical University of South Carolina

The Alzheimer’s Disease Neuroimaging Initiative (ADNI) captures longitudinal multivariate cognitive and imaging data of inherent complexity (between outcome correlations and within subject correlation of longitudinal measurements taken at subject-specific time intervals). However, a common method to analyze these types of data is to fit each outcome separately using a generalized linear mixed model (sGLMM), which ignores these between outcome correlations - a characteristic inherent to cognitive and imaging data. To address the limitations in existing methods, we propose a multivariate generalized linear mixed model (mGLMM) and assess its performance. We demonstrate a novel application of mGLMM in modeling multiple longitudinal outcomes (mix of cognitive and neuroimaging) from ADNI to study the difference in the trajectory of these outcomes by patient apolipoprotein epsilon 4 allele APOE(e4) and amyloid beta (A\(\beta\)) negativity status adjusting for age, sex, and education. We use eight outcomes that were measured longitudinally at baseline and every 4-6 months at 8 consecutive visits. To discern the optimal specification of the mGLMM, we vary its specification to include either independent, shared or separate random effects across outcomes, as well as either conditional independence or a dependence covariance structure within outcomes either homogeneously or heterogeneously specified across all eight outcomes. Model fitting was performed using a novel application of the GLMMIX procedure included in SAS 9.4, and model performance was determined on the basis of a variety of information criteria. Multiple imputation was performed to address missingness in these data, and parameter estimates were combined across imputations using standard procedures. We found that the mGLMM with shared random intercepts and a heterogeneous dependence covariance structure across outcomes fit our data best according to all available fit statistics. Furthermore, we found that the only exposure effect that was consistently significant across our mGLMM model variants was the main effect of A\(\beta\) negativity status. We failed to detect a significant interaction between A\(\beta\) negativity and APOE(e4) status. Thus, we conclude that the sharing of random effects across outcomes in the mGLMM framework is advantageous in modeling these ADNI data, and that A\(\beta\) negativity status (irrespective of time and APOE(e4) status) is significantly associated with ADNI patient cognitive status. Our proposed methods are applicable to a wide variety of multivariate longitudinal settings.

Bayesian Spatial Binary Regression for Label Fusion in Structural Neuroimaging

♦ Andrew Brown1, Chris McMahan1, Russell Shinohara2 and Kristin Linn2

1Clemson University
2University of Pennsylvania

Many analyses of neuroimaging data involve studying one or more regions of interest (ROIs) in a brain image. In order to do so, each ROI must first be identified. Since every brain is unique, the location, size, and shape of each ROI varies across subjects. Thus, each ROI in a brain image must either be manually identified or (semi-) automatically delineated, a task referred to as segmentation. Automatic segmentation often involves mapping a previously manually segmented image to a new brain image and propagating the labels to obtain an estimate of where each ROI is located in the new image. A more recent approach to this problem is to propagate labels from multiple manually segmented atlases and combine the results

62

2019 ICSA Applied Statistics Symposium, Raleigh, NC, June 9-12
using a process known as label fusion. To date, most label fusion algorithms either employ voting procedures or impose prior structure and subsequently find the maximum a posteriori estimator (i.e., the posterior mode) through optimization. We propose using a fully Bayesian spatial regression model for label fusion that facilitates direct incorporation of covariate information while making accessible the entire posterior distribution. We discuss the implementation of our model via Markov chain Monte Carlo and illustrate the procedure through both simulation and application to segmentation of the hippocampus, an anatomical structure known to be associated with Alzheimer’s disease.

**Weighted p-values for heterogeneous dependent data: an application to lesion symptom mapping**

1 Alexander Mclain, Joshua Habiger, Christopher Rorden and Julius Fridriksson

1 University of South Carolina
2 Oklahoma State University

ctliu@bu.edu

Recent advances in neuroimaging allow noninvasive measures of the structure and activity of the brain. Combined with behavioral tasks, these can reveal how specific areas of the brain are involved with cognitive functions. These studies result in data that can have thousands of observations per subject and require the use of thresholding procedures to control a global type I error rate. As a specific example, consider voxel-based lesion behavior mapping (VLSM) that relates areas of the brain with injury (e.g., from a traumatic brain injury, epilepsy, or stroke) to a cognitive outcome. VLSM studies cannot be designed to injure certain areas of the brain, thus naturally occurring lesions are used. The use of observational data results in hypotheses where, for example, lesion status (yes/no) is non-uniformly distributed across voxels with spatial dependence, resulting in heterogeneous power among tests. Applying standard multiple testing procedures results in testing hypotheses that are known a priori to have little to no power of being detected. Recent statistical methodology has proposed weighting hypotheses according to these prior beliefs. In this paper, we propose the use of p-value weighting to VLSM studies. We provide a significant theoretical result for a monotone minimum weight criteria which require minimum a priori power information, and address the issues dependence among hypotheses and confounding. Our methods are explored through simulation studies and an application to a study of stroke patients which examines regions of the brain associated with scores on the Western Aphasia Battery (WAB). The results demonstrate that the proposed methods can identify more significant hypotheses, and show how weights can be used to identify regions that are inconclusive due to lack of power.

**Session 2: Statistical Advances in Genetics, Genomics and Bioinformatics**

A unified method for rare variant analysis of gene-environment interactions

Elise Lim, Han Chen, Josee Dupuis and Ching-Ti Liu

1 Boston University
2 University of Texas Health Science Center at Houston

ctliu@bu.edu

Advanced technology in whole-genome sequencing has offered the opportunity to comprehensively investigate the genetic contribution, particularly rare variants, to complex traits. Several region-based tests have been developed to jointly model the marginal effect but methods to detect gene-environment (GE) interactions are underdeveloped. Identifying the modification effects of environmental factors on genetic risk poses a considerable challenge. To tackle this challenge, we develop a unified method to detect GE interactions of an array of rare variants with use of mixed effect model. The proposed method can accommodate either binary or continuous traits in related or unrelated samples. Under this model, genetic main effects, GE interactions, and sample relatedness are modeled as random effects. We adopt a kernel-based method to leverage the joint information across rare variants and implement variance component score tests to reduce the computational burden. Our simulation studies of continuous and binary traits show that the proposed method maintains correct type I error rates and high power under various scenarios, such as differing the direction of main genotype and GE interaction effects and the proportion of causal variants in the model for both continuous and binary traits. We apply our method to the Framingham Heart Study to test GE interaction of smoking on body mass index or overweight status. We replicate the CHRNA4 gene association reported in previous large consortium meta-analysis of single nucleotide polymorphism (SNP)-smoking interaction. Our proposed set-based GE test is computationally efficient and is applicable to both binary and continuous phenotypes, while appropriately accounting for familial or cryptic relatedness.

Stochastic Functional Linear Models for Gene-based Association Analysis of Quantitative Traits

1 Racong Fan, Shuqi Wang, Xiaohan Mei, Bingsong Zhang, Yue Han, Runqiu Wang and Fang Hong-Bin
2 Georgetown University Medical Center

rf740@georgetown.edu

Longitudinally measured phenotypes are important for exploring genetic and environmental factors that affect complex traits over time. Genetic analysis of multiple measures in longitudinal studies provides a valuable opportunity to understand genetic architecture and biological variations of complex diseases. Many genetic studies have been conducted in cohorts in which repeated measures on the trait of interest are collected on each participant over a period of time and sequence data are available. Such studies not only provide a more accurate assessment of disease condition, but enable us to investigate genes influencing on the trajectory of a trait and disease progression, which are likely to help reduce the remaining missing heritability of these traits. Although they are important, there is a paucity of statistical methods to analyze sequence data in longitudinal studies. In this paper, stochastic functional linear models are developed for temporal association analysis at gene levels to analyze sequence data and longitudinally measured quantitative traits. Functional data analysis techniques are utilized to reduce high dimensionality of sequence data and draw useful information. A variance-covariance structure is constructed to model the measurement variation and correlations of the traits based on the theory of stochastic processes. Spline models are used to estimate the time-dependent trajectory mean function. By intensive simulation studies, it is shown that the proposed stochastic models control type I errors well, and have higher power levels than those of the perturbation tests. We test and refine the models and related software using real data sets of Framingham Heart Study.

Gene-graph based imputation method for single-cell RNA sequencing data

Weimiao Wu, Yunqing Liu, Geoffrey Chupp, Xiting Yan and Zuo-heng Wang

Yale University
Session 3: Recent development in missing data and survey sampling problems

A Bayesian Growth Mixture Model for Complex Survey Data: Clustering Post-Disaster PTSD Trajectories

Rebecca Anthopolos\(^1\), Qixuan Chen\(^3\), Joseph Sedransk\(^2\), Mary Thompson\(^3\), Gang Meng\(^2\) and Sandro Galea\(^4\)

\(^1\)Columbia University
\(^2\)Joint Program in Survey Methodology, University of Maryland
\(^3\)University of Waterloo
\(^4\)Boston University

While growth mixture models (GMMs) have recently gained a lot of favor in population health research, research on GMMs for analyzing data from a complex sample survey is sparse. Existing literature uses a pseudo likelihood approach, which yields population averaged estimation and draws statistical inference using linearization or resampling methods to reflect the variability resulting from the sampling design. In this paper, we propose a Bayesian GMM using a full likelihood approach in which we incorporate sample design features either through covariates or through variance components. Estimation is subject-specific, and variance is modeled hierarchically to reflect stratification, clustering, and repeated measures. We further adjust for spatial correlations induced by proximity of sampling units. Our proposed model simultaneously estimates latent class-specific longitudinal trajectories and the effects of risk factors on predicting latent class membership. We develop an efficient Gibbs sampler for model fitting that includes only closed-form full conditional distributions, and show techniques for model selection and checking. We apply our proposed GMM to the Galveston Bay Recovery Study, a three-wave community-based panel survey, to characterize longitudinal trajectories of post-traumatic stress disorder among residents of Galveston and Chambers counties, Texas, in the aftermath of Hurricane Ike.

Identification of trans-eQTLs using mediation analysis with multiple mediators

Nayang Shan\(^1\), Zuoheng Wang\(^2\) and Lin Hou\(^3\)

\(^1\)Tsinghua University
\(^2\)Yale University

Identification of expression quantitative trait loci (eQTLs) advances our understanding of genetics and regulatory mechanisms of gene expression in various organisms. Previous studies suggest that trans-eQTLs may regulate expression of remote genes by altering the expression of nearby genes. Trans-association has been studied in the mediation analysis with a single mediator. However, prior applications with one mediator are prone to model misspecification due to correlations between genes. Motivated from the observation that trans-eQTLs are more likely to associate with more than one cis-gene than randomly selected SNPs in the GTEx dataset, we developed a computational method to identify trans-eQTLs that are mediated by multiple mediators. In simulation studies, multiple mediator analysis had increased power to detect mediated trans-eQTLs, especially in large samples. In the HapMap3 data, we identified 11 mediated trans-eQTLs that were not detected by the single mediator analysis in the combined samples of African populations. Moreover, the mediated trans-eQTLs in the HapMap3 samples are more likely to be trait-associated SNPs. Our approach has improved the power of detecting mediated trans-eQTLs and advanced knowledge of gene regulation.
from the same study center, considered the variability due to estimating propensity scores. Comparing the existing inverse propensity score weighting and adjustment methods and the proposed KW approach, Monte Carlo simulation studies were conducted and showed that the KW estimators reduced the bias and increases the efficiency of the estimated disease prevalence. The developed approach was further demonstrated using US National-Institutes-of-Health-American-Association-of-Retired-People cohort to estimate the nine-year all-cause mortality.

**Measures of the Degree of Departure from Ignorable Sample Selection for Non-Probability Samples**

*Brady T. West, Roderick Little, Phil Boonstra and Rebecca Andridge*

1University of Michigan  
2Ohio State University

bwest@umich.edu

With the current focus of survey researchers on “big data” that are not selected by probability sampling, measures of the degree of potential bias from non-random selection are sorely needed. Existing indices of the degree of departure from representative probability samples like the R-Indicator are based on functions of the propensity of inclusion, and are based on modeling the inclusion probability as a function of auxiliary variables. These methods are agnostic about the relationship between the inclusion probability and survey outcomes, which is a crucial feature of the problem. We propose a simple index of the degree of departure from ignorable selection for estimates of means that corrects this deficiency, called the Standardized Measure of Unadjusted Bias (SMUB). The index is based on a normal pattern-mixture model applied to the problem of sample selection, and is grounded in the model-based framework of non-ignorable selection, first proposed in the context of nonresponse by Rubin (1976 Biometrika). The methodology also provides for sensitivity analyses to adjust inferences for departures from ignorable selection, before and after adjustment for auxiliary variables. We also propose a simple index for estimates of proportions based on binary variables, and describe the underlying model for this index. Bayesian approaches to making inferences about the indices of selection bias will also be described. The proposed methods have been implemented in R, and are illustrated using data from the National Survey of Family Growth.

**Session 4: Advances in statistical methods for large-scale genomic studies**

**RNA-seq and Methyl-seq power calculation and design issue**

*Chien-Wei Lin* and *George Tseng*

1Medical College of Wisconsin  
2University of Pittsburgh

chlin@mcw.edu

diy@stat.oregonstate.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) and Methylation sequencing (Methyl-Seq) has gradually become a standard tool for global transcriptomic and epigenomic 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 the total budget and pre-specified unit experimental cost, the study design issue in RNA-Seq and Methyl-Seq is conceptually a more complex multi-dimensional constrained optimization problem rather than one-dimensional sample size calculation in traditional hypothesis setting. In this talk, we propose two statistical frameworks for each of applications, namely “RNASeqDesign” and “MethylSeqDesign”, to utilize pilot data for power calculation and study design of RNA-Seq and Methyl-Seq experiments. The approach is based on a 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 real applications to evaluate the performance.

**Estimating causal graphs from data with applications in genomics**

*Qing Zhou*

University of Carolina, Los Angeles

zhou@stat.ucla.edu

We develop a penalized likelihood estimation framework to learn the structure of Bayesian networks, represented as directed acyclic graphs, from different types of data. Our main contribution is a family of score-based structure learning algorithms, implemented in a new R package, sparsebn. When sufficient experimental intervention is available, our methods can construct causal networks among a large number of variables. We provide a comprehensive comparison of our approach to several popular structure learning methods on large graphs with hundreds to thousands of nodes, and show that our algorithms are more accurate and scale efficiently as the number of nodes increases. Several applications of our methods to large-scale genomics data will be discussed.

**Test-statistic correlation and data-row correlation**

*Bin Zhuo, Duo Jiang and Yanming Di*

Oregon State University

ydi@stat.oregonstate.edu

When a statistical test is repeatedly applied to rows of a data matrix—such as in differential-expression analysis of gene expression data, correlations among data rows will give rise to correlations among corresponding test statistic values. Correlations among test statistic values create many inferential challenges in false-discovery-rate control procedures, gene-set enrichment analysis, or other procedures aiming to summarize the collection of test results. To tackle these challenges, researchers sometimes will, explicitly or implicitly, use the correlations (e.g., as measured by the Pearson correlation coefficients) among the data rows to approximate the correlations among the corresponding test statistic values. We show that, however, such approximations are only valid under limited settings. We investigate the relationship between the correlation coefficient between a pair of test statistics (test-statistic correlation) and the correlation coefficient between the two corresponding data rows (data-row correlation). We derive an analytical formula for the test-statistic correlation as a function of the data-row correlation for a general class of test statistics: in particular, two-sample t-test is a special case. The analytical formula implies that the test-statistic correlation is generally weaker than the corresponding data-row correlation, and in general, the latter will not well approximate the former when the involved null hypotheses are false. We verify our analytical results through simulations.

A statistical method for identifying parallel and sequential me-
Double-Slicing Assisted Sufficient Dimension Reduction for High Dimensional Censored Data
Shanshan Ding$^1$, Wei Qian$^1$ and Lan Wang$^2$
$^1$University of Delaware
$^2$University of Minnesota

We propose a unified framework and an efficient algorithm for analyzing high-dimensional survival data under weak modeling assumptions. In particular, it imposes neither parametric distributional assumption nor linear regression assumption. It only assumes that the survival time $T$ depends on a high-dimensional covariate vector through low-dimensional linear combinations of covariates. The censoring time is allowed to be conditionally independent of the survival time given the covariates. This general framework includes many popular parametric and semiparametric survival regression models as special cases. The proposed algorithm produces a number of practically useful outputs with theoretical guarantees, including a consistent estimate of the sufficient dimension reduction subspace of, a uniformly consistent Kaplan-Meier type estimator of the conditional distribution function of $T$ and a consistent estimator of the conditional quantile survival time. Our asymptotic results significantly extend the classical theory of sufficient dimension reduction for censored data and the celebrated nonparametric Kaplan-Meier estimator to the setting where the number of covariates $p$ diverges exponentially fast with the sample size $n$. We demonstrate the promising performance of the proposed new estimators through simulations and a real data example.

Optimal tuning for divide-and-conquer kernel ridge regression with massive data
Ganggang Xu$^1$, Zuofeng Shang$^2$ and Guang Cheng$^3$
$^1$University of Miami
$^2$Indiana University-Purdue University Indianapolis
$^3$Purdue University

Tuning parameter selection is of critical importance for kernel ridge regression. To this date, data driven tuning method for divide-and-conquer kernel ridge regression (d-KRR) has been lacking in the literature, which limits the applicability of d-KRR for large data sets. In this paper, by modifying the Generalized Cross-validation (GCV, Wahba, 1990) score, we propose a distributed Generalized Cross-Validation (dGCV) as a data-driven tool for selecting the tuning parameters in d-KRR. Not only the proposed dGCV is computationally scalable for massive data sets, it is also shown, under mild conditions, to be asymptotically optimal in the sense that minimizing the dGCV score is equivalent to minimizing the true global conditional empirical loss of the averaged function estimator, extending the existing optimality results of GCV to the divide-and-conquer framework.

Integrative Linear Discriminant Analysis
Quefeng Li$^1$ and Lexin Li$^2$
$^1$University of North Carolina at Chapel Hill
$^2$University of California, Berkeley

Multiple types of data measured on a common set of subjects arise in many areas. Numerous empirical studies have found that integrative analysis of such data can result in better statistical performance in terms of prediction and feature selection. However, the advantages of integrative analysis have mostly been demonstrated empirically. In the context of two-class classification, we propose an integrative linear discriminant analysis method and establish a theoretical guarantee that it achieves a smaller classification error than running linear discriminant analysis on each data type individually. We address the issues of outliers and missing values, frequently encountered in integrative analysis, and illustrate our method through simulations and a neuroimaging study of Alzheimer’s disease.

Bayesian Deep Neural Networks for Nonlinear Variable Selection
Yan Sun, Qifan Song and Faming Liang
Purdue University

Variable selection plays an important role in data mining for high-dimensional nonlinear systems. However, most existing variable selection methods are either developed for linear systems or computationally infeasible, rendering imprecise or inefficient selection of relevant variables. We propose to employ a deep neural network to approximate the unknown nonlinear system and conduct variable selection under the Bayesian framework. We propose to draw samples from the posterior distribution using stochastic gradient Markov chain Monte Carlo, where both the hierarchical prior and shrinkage prior are considered. The proposed methods can execute very fast, establishing similar costs as optimization methods. However, they can perform much better than the optimization methods in identifying relevant variables for general high-dimensional nonlinear system.
Hierarchical community detection by recursive partitioning
Tianxi Li 1, Lihua Lei 2, Sharmodeep Bhattacharyya 3, Purnamrita Sarkar 4, Peter Bickel 5 and Elizaveta Levina 5
1 University of Virginia
2 University of California, Berkeley
3 Oregon State University
4 University of Texas at Austin
5 University of Michigan

Hierarchical community detection in networks is usually formulated as finding a single partition of the network into some “correct” number of communities. We argue that it is more interpretable and in some regimes more accurate to construct a hierarchical tree of communities instead. This can be done with a simple top-down recursive partitioning algorithm, starting with a single community and separating the nodes into two communities by spectral clustering repeatedly, until a stopping rule suggests there are no further communities. This class of algorithms is model-free, computationally efficient, and requires no tuning other than selecting a stopping rule. We show that there are regimes where this approach outperforms K-way spectral clustering, and propose a natural framework for analyzing the algorithm’s theoretical performance, the binary tree stochastic block model. Under this model, we prove that the algorithm correctly recovers the entire community tree under relatively mild assumptions. We also apply the algorithm to a dataset of statistics papers to construct a hierarchical tree of statistical research communities.

Finding the way for the Graph Matching Problem
Daniel Sussman 1 and Vince Lyzinski 2
1 Boston University
2 University of Massachusetts Amherst

Empirical likelihood (EL) methods, a nonparametric counterpart of maximum likelihood, provide the ability to estimate the distribution function and construct simultaneous confidence regions. However, the optimal EL-based confidence regions are often wide and difficult to compute. We argue is already outlined by Owen [2003] but computation is still left open with some optimization problems. We provide a computational algorithm to the optimal solution of two-sample likelihood ratio and construct confidence intervals based on it. Real data examples and simulation studies are provided to demonstrate its performance. The new methods are observed to have better finite sample performances than existing methods particularly when the censoring proportion is high. The new methods are also illustrated through a real data example.

Cost Regression Models with Censored Data
Xue Ding and Mai Zhou
University of Kentucky

Recent studies show that medical cost data can be heavily censored and highly skewed, which leads to more complex cost data analysis. Appropriate statistical analysis of cost data may lead to more cost-effective medical treatments, resulting in substantial cost savings. Therefore, it is important to identify factors that affect medical costs and to know how these factors influence medical costs. Median cost regression analysis is useful for this purpose. In this paper, we propose to use empirical likelihood (EL) methods based on influence function and jackknife techniques to construct confidence regions for regression parameters in median cost regression models with censored data. We further propose confidence intervals for the median cost with given covariates using the proposed EL-based confidence regions. Simulation studies are conducted to compare the proposed EL-based confidence regions with the existing normal approximation-based confidence regions in terms of coverage probabilities. The new EL-based methods are observed to have better finite sample performances than existing methods particularly when the censoring proportion is high. The new methods are also illustrated through a real data example.

A New Scope of Penalized Empirical Likelihood in High Dimensionality
Tong Tong Wu 1, Chengyong Tang 2 and Jinyuan Chang 2
1 University of Rochester
2 Temple University

Empirical likelihood (EL) methods, a nonparametric counterpart of likelihood methods, are appealing and effective, especially in conjunction with estimating equations through which useful informa-
tion can be adaptively and flexibly incorporated. It is also known in the literature that EL approaches encounter substantial difficulties when dealing with problems having high-dimensional model parameters and estimating equations. To answer the challenges, we propose a new penalized EL by applying two penalty functions respectively on the model parameters and the associated Lagrange multipliers in the optimizations of EL. Allowing both the dimensionalities of model parameters and estimating equations growing exponentially with the sample size, the estimator from our new penalized EL is sparse and consistent with asymptotically normally distributed nonzero components. We also design a statistical inference procedure for low-dimensional components of the model parameters, by linearly mapping the original estimating equations to a low-dimensional space. Nest coordinate descent algorithm, along with additional steps, can be used to tackle the well-known difficulties in EL computation, especially in high-dimensional settings. Simulation studies provide numeric evidence that the new penalized EL works well in high dimensionality. This method was applied to the TAAG data to examine multi-level factors related to the MVP levels over time for girls from adolescence into young adulthood.

Non-iterative Estimation for Semiparametric Models with Population-based Auxiliary Information

Fei Gao and Kwun Chuen Gary Chan
University of Washington
feigao@uw.edu

With the advancement in disease registries and surveillance data, population-based information on disease incidence, survival probability or other important biological characteristics become increasingly available. Such information can be leveraged in studies that collect detailed measurements but with smaller sample sizes. In contrast to recent proposals that formulate the additional information as constraints in optimization problems, we develop a general framework to construct simple estimators that update the usual regression estimators with some functionals of data that incorporate the additional information. We consider general settings which include nuisance parameters in the auxiliary information and semiparametric models with infinite dimensional parameters. Detailed examples of several important settings are provided.

Session 8: Functional data analysis: Foundations, Tools and Applications

Modeling Longitudinal Data on Riemannian Manifolds

Xiongtao Dai1, Zhenhua Lin2 and Hans-Georg Mueller2
1Iowa State University
2University of California, Davis
xda@iastate.edu

A manifold version of the principal analysis by conditional expectation (PACE) is proposed to represent sparsely observed longitudinal data that take values on a nonlinear Riemannian manifold. Typical examples of such manifold-valued data include longitudinal compositional data, as well as longitudinal shape trajectories located on a hypersphere. Compared to standard functional principal component analysis that is geared to Euclidean geometry, the proposed approach leads to improved trajectory recovery on non-linear manifolds in simulations. As an illustration, we apply the proposed method on longitudinal emotional well-being data for unemployed workers. An R implementation of our method is available on GitHub.

tidyfun: a New Framework for Representing and Working with Function-Valued Data

Fabian Scheipl1 and Jeff Goldsmith2
1Ludwig-Maximilians-Universität München
2Columbia University
ajg2202@cumc.columbia.edu

We present a new R package tidyfun for representing and working with function-valued data that presents a unified interface for dealing with regularly or irregularly observed function-valued data. The package follows the tidyverse design philosophy of R packages and is aimed at lowering the barrier of entry for analysts in order to quickly and painlessly analyse and interact with functional data and, specifically, datasets that contain both scalar and functional data or multiple types of functional data, potentially measured over different domains. We discuss the available feature set as well as forthcoming extensions and show some application examples.

Finding Biomarkers for Childhood Obesity Using Functional Data Analysis

Sarah Craig, Ana Kenney, Francesca Chiaromonte, Matthew Reimherr and Kateryna Makova
Pennsylvania State University
mreimherr@psu.edu

In this talk I will present recent work concerning the analysis of longitudinal childhood growth trajectories using functional data analysis. We explore both the microbiome and the genome for biomarkers that put children at greater risk for obesity. We discuss tools for both variable selection and parameter estimation when the outcome is functional and one has a large number of scalar predictors.

A Functional Mixed Model for Scalar on Function Regression with Application to Functional MRI Study

Wanying Ma1, Luo Xiao1, Bowen Liu1 and Martin Lindquist2
1North Carolina State University
2Johns Hopkins University
lxiao5@ncsu.edu

Motivated by a functional magnetic resonance imaging (fMRI) study, we propose a novel functional mixed model for scalar on function regression. The model extends the standard scalar on function regression for repeated outcomes by incorporating random subject-specific functional effects. Using functional principal component analysis, the new model can be reformulated as a mixed effects model and thus easily fit. A test is also proposed to assess the existence of the random subject-specific functional effects. We evaluate the performance of the model and test via a simulation study, as well as on data from the motivating fMRI study of thermal pain. The fMRI data application indicates significant subject-specific effects of the human brain hemodynamics related to pain and provides insights on how the effects might differ across subjects.

Session 9: Innovative methods for complex survival data and the applications

A Homoscedasticity in the Accelerated Failure Time Model

Lili Yu1 and Din Chen2
1Georgia Southern University
2University of North Carolina at Chapel Hill
dinchen@email.unc.edu

The semiparametric accelerated failure time (AFT) model is a popular linear model in survival analysis. Current research based on the AFT model assumed homoscedasticity of the survival data. Violation of this assumption has been shown to lead to inefficient and
Abstracts

Even unreliable estimation, and hence, misleading conclusions for survival data analysis. However, there is no valid statistical test in the literature that can be utilized to test this homoscedasticity assumption. This talk will discuss a novel quasi-likelihood ratio test for the homoscedasticity assumption in the AFT model. Simulation studies are conducted to show the satisfactory performance of this novel statistical test. A real dataset is used to demonstrate the application of this developed test.

Semiparametric regression analysis for composite endpoints subject to component-wise censoring

Guoqing Diao, Donglin Zeng, Chunlei Ke, Haijun Ma, Qi Jiang and Joseph Ibrahim

1 George Mason University
2 University of North Carolina at Chapel Hill
3 Amgen, Inc.

diao@gm.edu

Composite endpoints with censored data are commonly used as study outcomes in clinical trials. For example, progression-free survival is a widely used composite endpoint, with disease progression and death as the two components. Progression-free survival time is often defined as the time from randomization to the earlier occurrence of disease progression or death from any cause. The censoring times of the two components could be different for patients not experiencing the endpoint event. Conventional approaches, such as taking the minimum of the censoring times of the two components as the censoring time for progression-free survival time, may suffer from efficiency loss and could produce biased estimates of the treatment effect. We propose a new likelihood-based approach that decomposes the endpoints and models both the progression-free survival time and the time from disease progression to death. The censoring times for different components are distinguished. The approach makes full use of available information and provides a direct and improved estimate of the treatment effect on progression-free survival time. Simulations demonstrate that the proposed method outperforms several other approaches and is robust against various model misspecifications. An application to a prostate cancer clinical trial is provided.

Event-Specific Win Ratios and Testing with Terminal and Non-Terminal Events

Song Yang and James Troendle

National Institutes of Health

yangso@nih.gov

In clinical trials the primary outcome is often a composite one, defined as time to the first of two or more types of clinical events, such as cardiovascular death, a terminal event, and heart failure hospitalization, a non-terminal event. Thus if a patient experiences both types of events, then the terminal event after a non-terminal event does not contribute to the primary outcome, even though the terminal event is more important than the non-terminal event. If there are substantial number of patients who experience multiple events, the power of the test for treatment effect may be reduced due to omission of some of the available data. In the win ratio approach, priorities are given to the clinically more important events, and potentially all available data are used. However, the win ratio approach may have low power in detecting a treatment effect if the effect is predominantly on the non-terminal events. We propose event-specific win ratios obtained separately on the terminal and non-terminal events. These ratios can then be used to form global tests such as a linear combination test, the maximum test, or a Chi-square test. In simulations these tests often improve the power of the original win ratio test. Furthermore, when the terminal and non-terminal events experience differential treatment effects, the new tests often improve the power of the log-rank test for the composite outcome. Thus whether the treatment effect is primarily on the terminal events or the non-terminal events, the new tests based on the event-specific win ratios can be useful alternatives for testing treatment effect in clinical trials with time-to-event outcomes when different types of events are present. We illustrate the new tests with the primary outcome in the trial Aldosterone Antagonist Therapy for Adults With Heart Failure and Preserved Systolic Function (TOPCAT), where the new tests all reject the null hypothesis of no treatment effect while the composite outcome approach used in TOPCAT did not.

Session 10: Classification and modeling of multiple time series and functional data and applications

Development and Evaluation of Sensor Fire Prediction for Smart Appliances

Z. Q. John Lu, Amy Mensch and Anthony Hammins

National Institute of Standards and Technology

john.lu@nist.gov

Motivated by an ongoing project at NIST for developing an automated sensor ignition prevention system for smart cooktops, we discuss how data from such sensor networks should be analyzed and what metrics should be used in selecting sensors to prevent fire events. We find that novel graphical displays of multiple time series or functional data are important and help us select useful signals and features to extract as “signatures” of fire or non-fire events—learning representation needed for subsequent classification tasks. We also discuss the interesting issue of what role probability modeling such as logistic regression or support vector machines should play in such data-rich problems, which are characterized by complete- or quasi-separation of two events and with enough data to define classification rules empirically. Model-free metrics for performance evaluation of classifiers are also discussed.

Evaluating classification accuracy for modern learning approaches

Jialiang Li

National University of Singapore

stalij@nus.edu.sg

Deep learning neural network models such as multilayer perceptron (MLP) and convolutional neural network (CNN) are novel and attractive artificial intelligence computing tools. However, evaluation of the performance of these methods is not readily available for practitioners yet. We provide a tutorial for evaluating classification accuracy for various state-of-the-art learning approaches, including familiar shallow and deep learning methods. For quantitative response variables with more than two categories, many traditional accuracy measures such as sensitivity, specificity, and area under the receiver operating characteristic curve are not applicable and we have to consider their extensions properly. In this paper, a few important statistical concepts for multiclassification accuracy are reviewed and their utilities for various learning algorithms are demonstrated with real medical examples. We offer problem-based R code to illustrate how to perform these statistical computations step by step. We expect that such analysis tools will become more familiar to practitioners and receive broader applications in biostatistics.
A multi-group classification model for geographical origin discrimination of herbal medicines

Ying Zhu and Tuck Lee Tan
Nanyang Technological University
ying.zhu@nie.edu.sg

It is reported that for herbal medicines even from the same species, the quality and medicinal efficacy are quite different according to their growing conditions based on the geographical origins. Identification of herbal medicines from different origins remains a challenging issue due to their complex nature of mixture and different chemical compositions based on the cultivation areas. In this study Fourier transform infrared (FTIR) spectroscopy is used for the determination of geographic origins of herbal medicines. High dimensional FTIR spectroscopic data show highly correlated structure due to the complex system of spectroscopic data of herbal medicines. The interpretation of conventional discrimination methods becomes progressively complicated for multi-group situation. A multi-group classification model is developed for classifying the different geographical origins of herbal medicine samples by explaining class-specific differences and incorporating correlated spectral features. The proposed model yields more interpretable results in ing class-specific differences and incorporating correlated spectral different geographical origins of herbal medicines samples for clarifying the differences in samples from different origins.

A simple and adaptive two-sample test in high dimensions

Jin-Ting Zhang1, Jia Guo2, Bu Zhou3 and Ming-Yen Cheng4
1National University of Singapore
2Zhejiang University of Technology
3Zhejiang Gongshang University
4Hong Kong Baptist University

Test the equality of two means is a fundamental inference problem. For high-dimensional data, the Hotelling’s T-square test either performs poorly or becomes inapplicable. Several modifications have been proposed to address this issue. However, most of them are based on asymptotic normality of the null distributions of their test statistics which inevitably requires strong assumptions on the covariance matrix. We study this problem thoroughly and propose an L2-norm based test that works under mild conditions and even when there are fewer observations than the dimension. Specially, to cope with general non-normality of the null distribution we employ the Welch-Satterthwaite chi-square approximation. We derive a sharp upper bound on the approximation error and use it to justify that the chi-square approximation is preferred to normal approximation. Simple ratio-consistent estimators for the parameters in the chi-square approximation are given. Importantly, our test is adaptive to singularity or near singularity of the unknown covariance which is commonly seen in high dimensions and is the main cause of non-normality. The power of the proposed test is also investigated. Extensive simulation studies and an application show that our test outperforms several existing tests in terms of size control, and the powers are comparable when their sizes are comparable.

Session 11: Statistical Advances in High Dimensional Data Analysis

Co-Manifold Learning with Missing Data

Eric Chi1, Gal Mishne2 and Ronald Coifman2
1North Carolina State University
2Yale University

In this talk we introduce a new method for performing joint dimension reduction, or manifold learning, on the rows and columns of a data matrix. Our approach generalizes recent work on a convex formulation of the biclustering problem. Like convex biclustering, our co-manifold learning procedure possesses stability guarantees with respect to perturbations in the data. We illustrate how out method can identify coupled row and column geometries in simulated and real data examples.

Identifying noncoding RNAs and methylation biomarkers associated with tumor immunity

Xuefeng Wang and Xiaoning Yu
H. Lee Moffitt Cancer Center & Research Institute

Even with the great advances in immunotherapy in recent years, the response rate to existing immune checkpoint inhibitor therapy is very low in many cancer types. It is of high priority to discover new therapeutic targets and biomarkers that can better which patients can benefit from an immune checkpoint blocker. To date, there has not been a comprehensive analysis of the role of DNA methylation and noncoding RNAs in tumor immunity. This study is based on the publicly available gene expression and methylation data from The Cancer Genome Atlas cancer samples. We propose a high-dimensional-based screening method to identify candidate CpG and noncoding RNA biomarkers that are associated with a spectrum of immune outcomes, including cytolytic activity index, immune infiltration scores and T cell infiltration scores. We are able to identify a number of novel immune biomarkers, which are also associated with patient survival in many cancer types. Our analysis illustrates the potential of repurposing the RNAseq data and methylation data for a better understating of tumor-immune ecosystems.

LDSaRa: A Powerful Method for High Resolution Copy Number Variation Detection

Feifei Xiao and Xizhi Luo
University of South Carolina

Fast development of DNA sequencing technology offers a standard way of analyzing genomic variations, especially whole exome sequencing (WES), which directly identifies variants with respect to the coding regions. Among all the genomic structural variations, copy number variation (CNV) is a major component. Copy number analysis aims at comprehensively detecting and classifying CNVs, which provides scientists an alternative view of human genetic variations. However, most of the existing methods have two main issues. First, the widely used change-point based methods are developed with a strong assumption that observations of different loci are independent. However, this independence assumption is violated in the genetics perspective that the existence of linkage disequilibrium (LD) provides great insight into the modeling of correlation along the genome. Second, WES data is subjected to different types of biases which leads to false discoveries and low sensitivity that a proper normalization procedure is highly demanded. We previously developed an accurate and powerful CNV detection method with a screening and ranking algorithm (SaRa), which presented favorable feasibility and sensitivity. However, it assumes independent signals and mainly designed for the microarray experiments. To address these issues, we present LDSaRa as a novel improvement that integrates the LD information from external data or empirical estimates. Moreover, to extend SaRa to WES data, we propose LDSaRaseq, which includes two normalization options:
A Bayesian sparse latent factor model for identification of cancer subgroups with data integration

Zequn Sun, Brian Neelon and Dongjun Chung
Medical University of South Carolina

Identification of cancer subgroups is of critical importance for the development of precise therapeutic strategies for various types of cancer. The Cancer Genome Atlas (TCGA) have generated tremendous amount of high throughput genomic data, which profiles somatic mutation, copy number alteration, DNA methylation, gene expression for each patient. This large-scale cancer genomic data provides unprecedented opportunity to investigate cancer subgroups using integrative approaches based on multiple types of genomic data. In this presentation, I will discuss our recent work on a novel Bayesian sparse latent factor model for simultaneous identification of cancer subgroups (clustering) and key molecular features (variable selection) within a unified framework, based on a joint analysis of continuous (e.g., DNA methylation), binary (e.g., somatic mutation, copy number alteration), and count data (e.g., gene expression). In addition, by utilizing pathway (variable group) information, this approach does not only improve accuracy and robustness in identification of cancer subgroups and key molecular features, but also facilitates biological understanding of novel findings generated with this approach. Finally, in order to facilitate efficient posterior sampling, a heavy-tailed prior is specified for continuous data while alternative Gibbs samplers for logistic and negative binomial models are proposed based on Polya-Gamma mixtures of Normal densities for binary and count data. I will illustrate the proposed statistical model with simulation studies and its application to the TCGA data.

Session 12: Challenges in implementation of ICH E17

Experience of Reviewing MRCT Results and Expectations for ICH E17 in Taiwan

Guei-Feng Tsai
Statistical Team, CDE Taiwan
gftsai@cde.org.tw

Multi-Regional Clinical Trials (MRCTs) play a key role in the simultaneous global development of new drugs for pharmaceutical companies to avoid repeating trials and wasting resources. MRCTs with Asia/Asian subgroups have been increasingly utilized for regulatory submissions in Taiwan. In this talk, the speaker will share the regulatory experience of reviewing MRCT results and the expectations for the ICH E17 guideline in Taiwan.

Session 13: Historical Control in Clinical Trials

The Use of Historical Control in Clinical Trials in Rare Diseases

Chunqin Deng, Youlan Rao and Lisa Edwards
United Therapeutics
cdeng@unither.com

While RCT is the golden standard for clinical trials, in some situations, the single arm study without concurrent control become necessary. In studies without concurrent control, the efficacy and safety are demonstrated by comparing the study results with the historical control. The use of historical control is especially useful in clinical trials in rare diseases and in studies when the concurrent control is either unavailable or unethical. Two examples were used to demonstrate the usefulness of the historical control in clinical trials for supporting the product registrations: clinical program for Brineura’s approval for the treatment of Batten disease and the study of Irinotecan-temozolomide plus immunotherapy dinutuximab for the treatment of refractory or relapsed neuroblastoma in children. The issues related to the use of historical control in these studies will be discussed.

Statistical methodologies and challenges to properly unitize historical data

Rong Liu and Meredeh Ghadessi
1 Celgene
crliu@celgene.com

Several methods have been developed to analyze the historical data for epidemiology research, benchmarking and comparison of efficacy and safety profile of a drug or device. Some of these methods can be applied when only summary data from literature is available, and some of these methods require individual data. The challenges are to determine how much the historical controls (HCs) are similar and relevant to each other and to the current data, how much information can be borrowed from the HCs, what if the HCs are in conflict with each other or with the concurrent control data, and therefore how to incorporate the HCs in the analysis with minimum bias. In this talk, we will present both frequentist approach and Bayesian approach and will emphasize on Bayesian approaches. It is noteworthy that there is no one solution that works in all situations. The keys are simulations, and sensitivity analysis; simulation to create a benchmark to characterize each method for comparison to help with decision-making process, and sensitivity analysis to assess the robustness of the results. We will illustrate the methods with some examples.

Actual example of using historical control data for drug development

Kiichiro Toyoizumi
Shionogi, Inc.
kiichiro.toyoizumi@shionogi.co.jp

When implementing historical control data in a clinical trial, discussion points to consider are the source of the data, its quality, comparability of the data, the probability of erroneous conclusion and bias of a treatment effect. We have to consider the statistical methods to address these problems. Propensity score adjustment or Bayesian analysis can be candidates to address these problems. In this presentation, we will present some actual clinical trials submitted for regulatory approval which used historical control data using those methodologies.
Session 14: Using adaptive treatment strategies to give the right patient the right dose at the right time

Optimal Individualized Treatment Initiation Regime

Xin Chen, Rui Song, Jiajia Zhang, Liuquan Sun, and Wenbin Lu

1 Chinese Academy of Science
2 North Carolina State University
3 University of South Carolina
wlu4@ncsu.edu

When to initiate treatment on patients is an important problem in many medical studies such as AIDS and cancer. The goal of this article is to find an optimal individualized treatment initiation regime that determines the best treatment initiation time for individual patients based on their own characteristics. Different from existing optimal treatment regime for discrete choices of treatments, here the new challenges arise from the complicated missing mechanism in treatment initiation time data and the continuous treatment rule in terms of initiation time. To tackle these challenges, we proposed a new value function, which is constructed on restricted mean residual lifetime, to evaluate the performance of treatment initiation regime, and also developed a nonparametric kernel-based estimation for the value function which is shown to be consistent even when treatment initiation times are not completely observable and their distribution is unknown. We also studied the asymptotic properties of the resulting estimator in the decision rule and its associated value function estimator. In particular, the asymptotic distribution of the estimated value function is nonstandard, which follows a weighted chi-squared distribution. The finite-sample performance of the proposed estimation method is evaluated by simulation studies. An application to a breast cancer data is also provided for further illustration.

Subgroup-Specific Dose-Finding in Phase I Trials Based on Time to Toxicity

Peter Thall and Andrew Chapple

1 M.D. Anderson Cancer Center
2 Louisiana State University
rex@mdanderson.org

A new Bayesian design, Sub-TITE, is presented for phase I clinical trials with two or more pre-specified patient subgroups that does precision dose-finding based on time to toxicity. The design makes sequentially adaptive subgroup-specific decisions. Decisions are based on posterior quantities computed under a logistic regression model for the probability of toxicity within a fixed follow up period, as a function of dose and subgroup. Spike-and-slab priors are assumed for subgroup parameters, with latent subgroup variables included to allow different subgroups to be combined adaptively if they have similar estimated dose-toxicity curves. This framework is appropriate if clinicians have identified patient subgroups but are not certain whether they have different dose-toxicity curves. A simulation study shows that, when the dose-toxicity curves differ between all subgroups, Sub-TITE has superior performance compared to applying the TITE-CRM (2000) while ignoring subgroups, applying the TITE-CRM separately within subgroups, or using the maximum likelihood (ML) approach of Salter et al. (2015). When two or more subgroups are truly homogeneous but differ from other subgroups, Sub-TITE is greatly superior to all three alternative approaches. Practical guidelines and computer software are provided to facilitate application.

Improving Trial-based Treatment Strategies Using Evidence from Electronic Health Records

Peng Wu, Donglin Zeng, Haoda Fu, and Yuanjia Wang

1 Columbia University
2 University of North Carolina at Chapel Hill
3 Eli Lilly and Company
dzeng@email.unc.edu

Individualized treatment rules or recommendations (ITRs) tailor medical treatments according to patient-specific characteristics in order to maximize patient’s outcome. Data from randomized controlled trials (RCTs) are used to infer valid ITRs using statistical and machine learning methods. However, RCTs are usually conducted under specific inclusion/exclusion criteria, thus limiting the generalizability of ITRs to a broader real-world patient population. On the other hand, since patient’s electronic health records (EHRs) document treatment prescriptions in the real world, transferring information in EHRs to RCTs, if done appropriately, could potentially improve the performance of ITRs, in terms of precision and generalizability. In this work, we propose a new domain adaptation method to learn ITRs by incorporating evidence from EHRs. Specifically, we use EHRs to augment the feature space of the RCT and then learn the optimal ITRs stratifying by these features. We present theoretical justifications and conduct simulation studies. Finally, we apply our method to transfer information learned from EHRs of type 2 diabetes (T2D) patients to improve learning individualized insulin therapies from a RCT.

Session 15: Breakthroughs in the Analysis of Large-Scale and Complex Data

Efficient Signal Inclusion in Large-Scale Data Analysis

Jessie Jeng
North Carolina State University
xjjen@ncsu.edu

This work addresses the challenge of efficiently capturing a high proportion of true signals for subsequent data analyses when signals are not strong enough to be identified individually. We develop a new analytic framework focusing on false negative control under dependence. We propose the signal missing rate as a new measure to account for the variability of false negative proportion. Novel data-adaptive procedures are developed to control signal missing rate without incurring unnecessary false positives under dependence. The proposed methods are applied to GWAS on human heights to effectively remove irrelevant SNPs while retaining a high proportion of relevant SNPs for subsequent polygenic analysis.

Dynamic Scan Procedure for Detecting Rare-Variant Association Regions in Whole Genome Sequencing

Xihong Lin
Harvard School of Public Health
xlin@hsph.harvard.edu

Dynamic Scan Procedure for Detecting Rare-Variant Association Regions in Whole Genome Sequencing Studies Xihong Lin Department of Biostatistics and Department of Statistics Harvard University Whole genome sequencing (WGS) studies are being widely conducted to identify rare variants associated with human diseases and disease-related traits. Classical single-marker association analyses for rare variants have limited power, and variant-set based analyses are commonly used to analyze rare variants. However, existing variant-set based approaches need to pre-specify genetic regions for analysis, and hence are not directly applicable to WGS data due
to the large number of intergenic and intron regions that consist of a massive number of non-coding variants. The commonly used sliding window method requires pre-specifying fixed window sizes, which are often unknown as a priori, are difficult to specify in practice and are subject to limitations given genetic association region sizes are likely to vary across the genome and phenotypes. We propose a computationally efficient and dynamic scan statistic method (Scan the Genome (SCANG)) for analyzing WGS data that flexibly detects the sizes and the locations of rare-variants association regions without the need of specifying a prior fixed window size. The proposed method controls the genome-wise type I error rate and accounts for the linkage disequilibrium among genetic variants. It allows the detected rare variants association region sizes to vary across the genome. Through extensive simulated studies that consider a wide variety of scenarios, we show that SCANG substantially outperforms several alternative rare-variant association detection methods while controlling for the genome-wise type I error rates. We illustrate SCANG by analyzing the WGS lipids data from the Atherosclerosis Risk in Communities (ARIC) study.

Analysis of Data in Cones

Lingsong Zhang

Purdue University

Lingsong@purdue.edu

Complex data arise more often in applications such as images, genomics and many others. Traditional data were analyzed based on theoretical assumptions of data lie in Euclidean space. Recent years many new data types are within restricted space or sets, and require a new set of theory and methodology to analyze it. In this talk, we will focus on two types of data that lies in cones, and propose a generalized principal component type of tools to reveal underlying structure (or hidden factors) within such data. The approach naturally forms nested structure and thus is suitable for future investigation of optimal dimension. Application of this method such as diffusion tensor images will be shown in this talk as well.

Two component semiparametric density mixture with a known component: estimation and applications

Zhou Shen1 and ♦Michael Levine2

1J.P. Morgan
2Purdue University

mlevins@purdue.edu

We consider a semiparametric mixture of two univariate density functions where one of them is known while the other function are unknown. Such mixtures have a history of application to the problem of detecting differentially expressed genes under two or more conditions in microarray data. Until now, some additional knowledge about the unknown component (e.g. the fact that it belongs to a location family) has been assumed. As opposed to this approach, we do not assume any additional structure on the unknown density function. We derive a new sufficient identifiability condition for this model. We also suggest a novel approach to estimation of this model that is based on an idea of applying a maximum smoothed likelihood to what would otherwise have been an ill-posed problem. We introduce an iterative MM (Majorization-Minimization) algorithm that estimates all of the model parameters. We establish that the algorithm possesses a descent property with respect to a log-likelihood objective functional and prove that the algorithm, indeed, converges. Finally, we also illustrate the performance of our algorithm in a simulation study and using a real dataset.

Semiparametric inference for merged data from multiple sources

Takumi Saegusa

University of Maryland, College Park

tsaegusa@math.umd.edu

We study semiparametric inference for merged data from multiple data sources. A setting we consider is characterized by (1) duplication of the same units in multiple samples, (2) unidentified duplication across samples, (3) dependence due to finite population sampling. Applications include data synthesis of clinical trials, epidemiological studies, disease registries and health surveys. Our approach is to view two-phase stratified sampling as data integration from non-overlapping sources whereby theory and methods from the analysis of stratified samples continue to prove useful. Main results are the extension of empirical process theory to biased and dependent samples with duplication. Specifically we develop the uniform law of large numbers and uniform central limit theorem with applications to general theorems for consistency, rates of convergence and asymptotic normality in the context of data integration. Our results are illustrated with simulation studies and a real data example using the Cox proportional hazards model.

Variable selection for multiple types of high-dimensional features with missing data

♦Kin Yau Wong1, Donglin Zeng2 and Danyu Lin2

1The Hong Kong Polytechnic University
2University of North Carolina at Chapel Hill

kin-yau.wong@polyu.edu.hk

danlin@unc.edu

danyulin@unc.edu

In biological, clinical, and epidemiological studies, one often encounters high-dimensional data of multiple types or modalities. Because complete data from all modalities are rarely available for all study subjects, such multi-modality data often exhibit a substantial amount of missingness. A common approach to handle incomplete (high-dimensional) data is to obtain a complete data set using, for example, listwise deletion or single imputation and apply conventional statistical methods on the complete data set. This approach, however, is inefficient and may even be invalid. In this presentation, we consider variable selection with multiple types of potentially missing features. We propose a latent variable model to characterize the relationships across and within feature types and to infer missing values from observed data. We develop a penalized-likelihood approach for variable selection and estimation and de-
vise an efficient expectation-maximization algorithm for its implementation. Finally, we demonstrate the satisfactory performance of the proposed methods through simulation studies and provide an application to a motivating multi-platform genomics study that involves high-dimensional features with missing data. By adopting a likelihood-based approach with a low-dimensional factor model for the incomplete variables, the proposed methods tend to be more efficient and accommodate more general missing-data patterns and mechanisms than existing methods.

**Case-cohort Studies with Multiple Interval-censored Disease Outcomes**

* Qingning Zhou¹, Jianwen Cai² and Haibo Zhou³
¹University of North Carolina at Charlotte
²University of North Carolina at Chapel Hill
³qzhou@uncc.edu

Interval-censored failure time data commonly arise in epidemiological and biomedical studies where the occurrence of an event or a disease is determined through periodic examinations. Subject to interval-censoring, available information on the failure time can be quite limited. Cost-effective sampling designs are desirable to enhance the study power, especially when the disease rate is low and the covariates are expensive to obtain. The existing designs were mainly developed for studies concerning a single disease outcome. In many applications, it is of interest to compare the effects of a risk factor on multiple diseases. We formulate the case-cohort design with multiple interval-censored disease outcomes and generalize it to nonrare diseases where only a portion of diseased subjects are sampled. For inference, we develop a marginal sieve weighted likelihood approach. We assume that the failure times marginally follow the proportional hazards model, and construct a pseudo-likelihood function under working independence assumption. We consider two types of weights to account for the sampling bias. We adopt a sieve approach with Bernstein polynomials to handle the unknown baseline functions. We employ a weighted bootstrap procedure to obtain a robust variance estimate that is consistent regardless of the dependence structure between failure times. The proposed design and method are examined through simulation studies and illustrated with a dataset on diabetes and hypertension from the Atherosclerosis Risk in Communities (ARIC) study.

**Session 17: Different quantities of interest in observational studies**

**Statistical Inference on Shape and Size Indexes for Recurrent Event Processes**

* Yifei Sun¹, Steven Chiou² and Chiung-Yu Huang³
¹Columbia University
²University of Texas at Dallas
³University of California, San Francisco
ys3072@cunmc.columbia.edu

Single index models have gained increased popularity in time-to-event analysis owing to its flexibility as well as interpretable covariate effects. Recurrent event data has been studied in various fields such as reliability, medicine, social sciences and economics. Motivated by the research interest in understanding the covariate effect on a recurrent event process, we propose to use a shape index and a size index to provide a comprehensive view of the event risk. The proposed model possesses the interpretability of parameters via monotonic constraints on the link functions. In the presence of censoring, the two unknown link functions in the shape and size components create substantial analytical challenges. We develop a two-step rank-based estimation procedure. The proposed estimators are asymptotically normal with a root-n convergence rate. Moreover, we develop testing procedures for the null hypothesis that the shape and size function does not depend on covariates. Simulation studies and a data example are presented to illustrate the proposed methodology.

**Parsimonious Regressions for Repeated Measure Analysis**

* Dennis Cook¹, Liliana Forzani and Yichen Cai²
¹University of Minnesota
¹lychenang@umn.edu
²University of Wisconsin-Madison
²liux3771@umn.edu

Longitudinal data with repeated measures frequently arise in various disciplines. The standard methods typically impose a mean outcome model as a function of individual features, time and their interactions. However, the validity of the estimators relies on the correct specifications of the time dependency. The envelope method is recently proposed as a sufficient dimension reduction (SDR) method in multivariate regressions. In this paper, we demonstrate the use of the envelope method as a new parsimonious regression method for repeated measures analysis, where the specification of the underlying pattern of time trend is not required by the model. We found that if there is enough prior information to support the specification of the functional dependency of the mean outcome on time and if the dimension of the prespecified functional form is low, then the standard method is advantageous as an efficient and unbiased estimator. Otherwise, the envelope method is appealing as a more robust and potentially efficient parsimonious regression method in repeated measure analysis. We compare the performance of the envelope estimators with the existing estimators in a simulation study and in an application to the China Health and Nutrition Survey.

**When Wilcoxon-Mann-Whitney Meets Non-Compliance**

* Lu Mao
University of Wisconsin-Madison
lmao@biostat.wisc.edu

The nonparametric, rank-based Wilcoxon-Mann-Whitney (WMW) test is a popular alternative to the standard t-test for its robustness and potential efficiency. In addition, the WMW test statistic also provides an easily interpretable measure of effect size, in terms of probabilistic shift. Cast in the Rubin Causal Model, the probabilistic (shift) index has received considerable attention under randomized trials and under non-randomized trials with measured confounders. While non-compliance with unknown confounders is a common presence in randomized trials, little is known about the proper use or adaptation of the WMW statistics in such cases. Following the seminal work of Angrist, Imbens, and Rubin (1996) on the local average treatment effect on the compliers, we seek to establish a parallel, instrumental-variable-based framework for WMW statistics with an insistence on causal interpretability. Three questions are of interest: 1. In the presence of non-compliance, what constitutes a local probabilistic index on the compliers? 2. How do we construct an estimator for the local probabilistic index? 3. How does the test based on this estimator compare with the standard intent-to-treat WMW test in terms of power? We will show that these questions can be answered using the idea of principal stratification. The resultant procedures are applied to a job training study as an illustration.

**Solar Flare Predictions with Machine Learning**

* Yang Chen, Ward Manchester, Alfred Hero, Gabor Toth and Tamas Gombosi
University of Michigan
ychenang@umich.edu

The solar flare dataset comprises 53 characteristics of solar flares obtained by the Solar and Heliospheric Observatory (SOHO) spacecraft, along with an indicator of whether the flare was a C-class flare or greater. Using a stepwise approach, we develop a logistic regression model to discriminate between C-class or greater flares and all other events. We then use random forests to identify important features for flare prediction, and use gradient boosting to predict the occurrence of C-class or greater flares. Finally, we optimize our models for a job training study as an illustration.
Over the space age, we have accumulated extensive knowledge of the regions of space surrounding the Earth and the Sun, and the governing physical processes controlling space weather in these regions. However, this knowledge has not been translated into an operational forecast capability. By combining our expertise in space weather modeling and data science/machine learning, we can not only address the "holy grail" of space weather prediction and extend the forecast horizon from minutes to days, but also transition the results to space weather operations. The current space weather predictive capabilities are either short term and/or not accurate and reliable. We initiated a research program that will (hopefully) answer these questions using a unique combination of modeling, computation, and massive amounts of space weather data collected by satellites that will be used to train state-of-the-art machine learning algorithms.

Session 18: Quantitative decision making in early clinical development

An Evidence Based Approach for Phase II Proof-of-Concept Trial
Satrajit Roychoudhury
Pfizer, Inc.
satrajit.roychoudhury@pfizer.com

Proof-of-concept (PoC) trials play a critical role in the clinical development of an experimental therapy. Typically, these trials offer the first read-out of efficacy and lead to one of the following decisions: consider further development (GO), stop development (NO-GO), or seek further information. To achieve that goal, statistical significance is assessed but usually fails to produce efficient decision in absence of clinically-relevant effect estimates. To palliate this, we propose a dual-criterion design which formally combines a statistical and clinical criterion. The dual-criterion design requires two inputs; a null hypothesis (i.e., no effect) and a decision value (i.e., minimum effect estimate). Unlike the standard design, with statistical significance as the sole success criterion, the decision value is explicit while the study power is implicit. Sample size determination in dual-criterion design requires special attention with regard to operating characteristics (i.e., error rates) and implied study outcomes. We successfully applied the dual-criterion design in oncology Phase II trials with binary and time-to-event endpoints. The evaluation covered a characterization of the decision criteria, sample size, as well as data scenarios and operating characteristics. Yet, despite their apparent simplicity, those designs can be conceptually challenging especially in terms of implementation and communication between the statistician and the trial design team. When properly understood and well executed, dual-criterion design based on statistical significance and clinically relevant effect size improves evidence-based, quantitative decision-making in early phase PoC trials.

Considerations for Basket Trial Design under Multisource Exchangeability Assumptions
*Michael Kane1, Alex Kaizer2, Nan Chen3 and Brian Hobbs4
1Yale University
2University of Colorado Denver
3University of Texas MD Anderson Cancer Center
4The Cleveland Clinic
michael.kane@yale.edu

Multisource Exchangeability Models (MEM) quantify the extent to which power can be borrowed across arms in a basket trial. This class of trial design is extremely useful in immunotherapy studies where the treatment acts on an underlying mechanism across multiple indications. In these cases, the pooling rare cancer types seeing poor enrollment may be pooled with similarly responding arms thereby reducing the required sample size and potentially reducing the time needed to accrue patients. In this talk we will review the MEM methodology and examine the behavior of the trials under a variety of realistic scenarios using real-world trial data.

Quantitative decision making in proof of mechanism study
Weidong Zhang
Pfizer, Inc.
weidong.zhang2@pfizer.com

In early clinical development, proof of mechanism (PoM) study is of great importance in assessment of pharmacology, specifically three pillars of drug development. It was shown that successful demonstration of PoM could significantly improve probability of success of PoC studies. However, declarations of PoM studies frequently rely on visual assessment of the data, and often, historical data are not effectively incorporated into the decision process. Hence the decision making process of PoM can be subjective and may pose higher risk of false decisions. In this presentation, we developed a Bayesian approach that can be used to objectively evaluate pharmacology using biomarkers. Using this approach, clinicians are able to take into account all information including data collected from current study and data from previous studies, by which they are given a few metrics to measure drug modulation effect and statistical confidence. Objective assessment can be made using this approach and false decisions can be significantly minimized.

Session 19: Causal inference and missing data analysis: identification and estimation

Causal inference with confounders missing not at random
*Shu Yang1, Linbo Wang2 and Peng Ding3
1North Carolina State University
2University of Toronto
3University of California, Berkeley
syang24@ncsu.edu

It is important to draw causal inference from unconfounded observational studies, which, however, becomes challenging if the confounders have missing values. Generally, causal effects are not identifiable if the confounders are missing not at random. We propose a novel framework to nonparametrically identify causal effects with confounders subject to an outcome independent missingness, that is, the missing data mechanism is independent of the outcome, given the treatment and possibly missing confounders. We then propose a nonparametric two stage least squares estimation and a parametric estimation for causal effects.

The Blessings of Multiple Causes
*Yixin Wang and David Blei
Columbia University
yixin.wang@columbia.edu

Causal inference from observational data often assumes "strong ignorability," that all confounders are observed. This assumption is standard yet untestable. However, many scientific studies involve multiple causes, different variables whose effects are simultaneously of interest. We propose the deconfounder, an algorithm that combines unsupervised machine learning and predictive model checking to perform causal inference in multiple-cause settings.
The deconfounder infers a latent variable as a substitute for unobserved confounders and then uses that substitute to perform causal inference. We develop theory for when the deconfounder leads to unbiased causal estimates, and show that it requires weaker assumptions than classical causal inference. We analyze its performance in three types of studies: semi-simulated data around smoking and lung cancer, semi-simulated data around genomewide association studies, and a real dataset about actors and movie revenue. The deconfounder provides a checkable approach to estimating closer-to-truth causal effects.

**Analysis of clustered data with cluster-specific nonignorable nonresponse**

*Danhyang Lee and Jae-Kwang Kim*

Iowa State University
[danhyang@iastate.edu](mailto:danhyang@iastate.edu)

Random effect models are popular in handling clustered data. If the study variable is subject to missingness and the response mechanism depends on the unobservable cluster-specific effect, then the assumption of missing at random (Rubin, 1976) does not hold and the classical analysis assuming MAR can lead to biased results. Such a nonresponse mechanism is called cluster-specific nonignorable (CSNI) nonresponse. We propose a new approach of parameter estimation for random effect models under CSNI nonresponse. Our method is based on the idea of imputing missing quantities in the full data EM algorithm with their unbiased predictors. Then, we propose a modified version of the EM algorithm to adjust for any biases due to the imputation. The proposed method does not require a correct model specification for the response mechanism. A limited simulation study is presented to show the performance of the proposed method.

**Recent Advances and Best Practices in Comparative Analyses using Real World Data**

*Douglas Faries*

Eli Lilly and Company
[dl.faries@lilly.com](mailto:dl.faries@lilly.com)

Real world data research in the pharmaceutical industry continues to expand in scope and impact. Traditional real world claims database and prospective observational studies for epidemiology insights or access and reimbursement support is now expanding into discussions of linking data sources and regulatory use of RWE. While there is growing use, there are issues regarding data quality, study design and analysis that need addressing in order for comparative real world data analyses to provide reliable information for decision making. In this talk, we will discuss real-world evidence research in Pharma as well as methodological issues and innovations that could improve the operating characteristics of real world research. Discussions will include research designs, such as the value and challenges of pragmatic trials, as well as methodological improvements for bias control and addressing unmeasured confounding.

**Session 20: Design and analysis of complex survival data**

**Joint modeling of survival and longitudinal quality of life data with informative censoring**

*Zhiguo Li, Janaka Perugaswaththeliyan and Lihui Zhao*

1University of Florida
[zhiguo.li@duke.edu](mailto:zhiguo.li@duke.edu)

2Northwestern University
[zhiguo.li@duke.edu](mailto:zhiguo.li@duke.edu)

Palliative medicine is an interdisciplinary specialty focusing on improving quality of life (QOL) for patients with serious illness and their families. Palliative care programs are widely available or under development at US hospitals. In palliative care studies, often longitudinal QOL and survival data are highly correlated which, in the face of censoring, makes it challenging to properly analyze and interpret terminal QOL trend. Informative dropout in the study add another level of complication of the problem. To address these issues, we propose a novel statistical approach to jointly model the terminal trend of QOL and survival data accounting for informative dropout. We assess the model through simulation and application to establish a novel modeling approach that could be applied in future palliative care treatment research trials.

**Sample size calculation for cluster randomization trials with a time-to-event endpoint**

*Sin-Ho Jung and Jianghao Li*

Duke University
[sinho.jung@duke.edu](mailto:sinho.jung@duke.edu)

Cluster randomization trials randomize groups (called clusters) of subjects (called subunits) between intervention arms, and observations are collected from each subject. In this case, subunits within each cluster share common frailties, so that the observations from subunits of each cluster tend to be correlated. Oftentimes, the outcome of a cluster randomization trial is a time-to-event endpoint with censoring. In this paper, we propose a closed form sample size formula for rank tests to compare the marginal survival distributions between intervention arms under cluster randomization with possibly variable cluster sizes. Extensive simulation studies were conducted to evaluate the performance of our sample size formula under various design settings. Real study examples are taken to demonstrate our method.

**Sample Size Calculation for Studies with Grouped Survival Data**

*Zhiguo Li, Xiaofei Wang, Yuan Wu and Kouros Owzar*

Duke University
[zhiguo.li@duke.edu](mailto:zhiguo.li@duke.edu)

Grouped survival data arise often in studies where the disease status is assessed at regular visits to clinic. The time to the event of interest can only be determined to be between two adjacent visits or is right censored at one visit. In data analysis, replacing the survival time with the endpoint or midpoint of the grouping interval leads to biased estimators of the effect size in group comparisons. Prentice and Gloeckler developed a maximum likelihood estimator for the proportional hazards model with grouped survival data and the method has been widely applied. Previous work on sample size calculation for designing studies with grouped data is either based on the exponential distribution assumption or approximation of variance under the alternative with variance under the null. Motivated by studies in HIV trials, cancer trials as well as in vitro experiments to study drug toxicity, we develop a sample size formula for studies with grouped survival endpoints that use Prentice and Gloeckler’s method for comparing two arms under the proportional hazards assumption. We do not impose any distributional assumptions, nor do we use any approximation of variance of the test statistic. The sample size formula only requires estimates of the hazard ratio and survival probabilities of the event time of interest and the censoring time at the endpoints of the grouping intervals for one of the two arms. The formula is shown to perform well in a simulation study and its application is illustrated in the three motivating examples.

**SampleN: an R-shiny APP to calculate sample size, including subgroup analysis for survival outcomes**

*Qian Wu and Isaac Jenkins*
SampleN: an open-source R-shiny APP to calculate sample size and protocol writing guidance in clinical trial design, including subgroup analysis for survival outcomes. Sample size estimation is an important step in clinical trial design that often needs thorough statistical knowledge and protocols writing experience. The current sample size calculation software, e.g., PASS, EAST, nQuery, are expensive and not available on MAC. The available online free tools included limited functions and not designed to help protocol writing. In order to provide an accurate and reliable tool, we developed and implemented commonly used sample size functions into an open-source, user-friendly, web-based, R-shiny APP named “SampleN”. SampleN is a free tool, for both statisticians and scientists. It has useful features: (1) sample size calculation for continuous, binary and survival endpoints (additional links for negative binomial) (2) template for statistical section in both clinical trial protocol (phase I/II/III) and grant writing (power analysis) (3) source codes in R based on user’s interface, free to download. SampleN has allowed many scientists the opportunity to calculate accurate sample sizes for experimental designs, making this a good tool to professionals in different fields. To illustrate the use of SampleN, a subgroup analysis with survival outcomes by using high-dimensional biomarker data were presented.

Proportional subdistribution hazards regression for competing risks data under case-cohort studies

Soyoung Kim, Yayun Xu, Mei-Jie Zhang and Kwang Woo Ahn
akim@mcw.edu
The case-cohort study design is widely used to reduce cost when collecting expensive covariates in large cohort studies with survival or competing risks outcomes. The subdistribution hazards model directly evaluates the effect of a covariate on the cumulative incidence function under the non-covariate-dependent censoring assumption. However, the non-covariate-dependent censoring assumption is often violated in many biomedical studies. In this paper, we propose a proportional subdistribution hazards model for case-cohort studies with stratified data with covariate-adjusted censoring weight. We further propose an efficient estimator when extra information from the other causes is available under case-cohort studies. The proposed estimators are shown to be consistent and asymptotically normal. Simulation studies show (i) the proposed estimator is unbiased when the censoring distribution depends on covariates; and (ii) the proposed efficient estimator gains estimation efficiency. We analyze a bone marrow transplant data set and a coronary heart disease data set using the proposed method.

Session 21: Nonconvex Optimization and Panel Data Analysis

Nonconvex Regularized Robust Regression with Oracle Properties in Polynomial Time

Wenxin Zhou1, Xiaosou Pan1 and Qiang Sun2
1University of Carolina, San Diego
2University of Toronto
wezx4@ucsd.edu
In this talk, we discuss the tradeoff between optimization complexity and statistical error for penalized robust regression. In the presence of heavy-tailed errors, we show that adaptive Huber regression with nonconvex regularization yields statistically optimal estimators that satisfy oracle properties as if the true active set were known. Computationally, we need at most $O(\log s + \log \log d)$ convex relaxations to reach such oracle estimators, where $s$ and $d$ denote the sparsity and ambient dimension, respectively. Numerical studies lend strong support to our methodology and theory.

On the Global Convergence of Policy Optimization in Deep Reinforcement Learning

Zhaoran Wang1, Qi Cai, Jason Lee, Boyi Liu and Zhurong Yang
Northwestern University
zhaoranwang@gmail.com
Policy optimization (with neural networks as actor and critic) is the workhorse behind the success of deep reinforcement learning. However, its global convergence remains less understood, even in classical settings with linear function approximators. In this talk, I will show that coupled with neural networks, a variant of proximal/trust-region policy optimization (PPO/TRPO) globally converges to the optimal policy. In particular, I will illustrate how the overparameterization of neural networks enable us to establish strong guarantees.

Structure Learning in Panel Data Analysis

Yuan Ke1, Jialiang Li2 and Wenyang Zhang3
1University of Georgia
2National University of Singapore
3University of York
yuan.ke@uga.edu
Panel data analysis is an important topic in statistics and econometrics. In panel data analysis, it is very common to assume the impact of a covariate on the response variable is the same across all individuals. Whilst the modelling based on this assumption is reasonable when only the global picture is of interest, in general, it may miss some individual/subgroup attributes of the impact. In this paper we propose a data driven approach to identify the groups in panel data with interactive effects caused by latent variables. It is assumed that the impact of a covariate is the same within each group, but different between the groups. An EM based algorithm is proposed to estimate the unknown parameters, and a binary segmentation based algorithm is proposed to do the grouping. Asymptotic properties are established to justify the proposed estimation, grouping method, and the modelling idea. Simulation studies are also conducted to compare the proposed method with the existing ones, and the results obtained favor our method. Finally, the proposed method is applied to analyze a data set about income dynamics, which leads to some interesting findings.

Session 22: Recent advances on high dimensional statistical learning and algorithm

First-order Newton-type Estimator for Distributed Estimation and Inference

Xi Chen1, Weidong Liu2 and Yichen Zhang3
1New York University
2Shanghai Jiao Tong University
xichen3@stern.nyu.edu
We consider a distributed estimation and inference for a general statistical problem with a convex loss that can be non-differentiable. We develop a new multi-round distributed estimation procedure to some interesting findings.

2019 ICSA Applied Statistics Symposium, Raleigh, NC, June 9-12 77
of estimating the Hessian matrix that usually requires the second-order differentiability of the loss, our estimator, called First-order Newton-type Estimator (FONE), directly estimates the vector of interest as a whole and is applicable to non-differentiable losses. Moreover, our method kills two birds with one stone. In the limiting distribution result, it turns out that the key term in the limiting covariance also has a similar form of the multiplication of the inverse population Hessian matrix and a given vector, which can also be estimated by FONE. The proposed FONE has many other potential applications to statistical estimation problems such as linear discriminant analysis (LDA).

Joint association and classification analysis of multi-view data
Yanjeng Zhang and Irina Gaynanova
Texas A&M University
irinag@stat.tamu.edu

Multi-view data, that is matched sets of measurements on the same subjects, have become increasingly common with technological advances in genomics and other fields. Often, the subjects are separated into known classes, and it is of interest to find associations between the views that are related to the class membership. Existing classification methods can either be applied to each view separately, or to the concatenated matrix of all views without taking into account between-views associations. On the other hand, existing association methods can not directly incorporate class information. In this work we propose a framework for Joint Association and Classification Analysis of multi-view data (JACA). We support the methodology with theoretical guarantees for estimation consistency in high-dimensional settings, and numerical comparisons with existing methods. In addition to joint learning framework, a distinct advantage of our approach is its ability to use partial information: it can be applied both in the settings with missing class labels, and in the settings with missing subsets of views. We apply JACA to colorectal cancer data from The Cancer Genome Atlas project, and quantify the association between RNAseq and miRNA views with respect to consensus molecular subtypes of colorectal cancer.

Ensemble estimation and variable selection with semiparametric regression models
Sunyoung Shin1, Yufeng Liu2, Stephen Cole3 and Jason Fine2
1University of Texas at Dallas
2University of North Carolina at Chapel Hill
sunyoung.shin@utdallas.edu

Scenarios are considered in which the likelihood function for a semiparametric regression model factors into separate components, with an efficient estimator of the regression parameter available for each component. An optimal weighted combination of the component estimators, named an ensemble estimator, may be employed as an overall estimate of the regression parameter, and may be fully efficient under uncorrelatedness conditions. This approach is useful when the full likelihood function is difficult to maximize but the components are easy to maximize. As a motivating example, we consider proportional hazards regression with prospective doubly-censored data, in which the likelihood factors into a current status data likelihood and a left-truncated right-censored data likelihood. Variable selection is important in such regression modelling but the applicability of existing techniques is unclear in the ensemble approach. We propose ensemble variable selection using the least squares approximation technique on the unpenalized ensemble estimator, followed by ensemble re-estimation under the selected model. The resulting estimator has the oracle property such that the set of nonzero parameters is successfully recovered and the semi-parametric efficiency bound is achieved for this parameter set. Simulations show that the proposed method performs well relative to alternative approaches. Analysis of the multicenter AIDS cohort study illustrates the practical utility of the method.

Recovery of Sums of Sparse and Dense Signals by Incorporating Graphical Structure Among Predictors
Yiyun Luo and Yufeng Liu
University of North Carolina at Chapel Hill
yiyun@stat.unc.edu

With the abundance of high dimensional data, sparse regularized techniques are very popular in practice due to the built-in sparsity of the solutions. However, the success of these sparse methods heavily relies on the sparsity assumption of the underlying model. For models where the sparsity assumption fails, the performance of these sparse methods can be unsatisfactory and misleading. To relax the sparsity assumption, we consider the perturbed linear model, where the signal is given by the sum of sparse and dense signals. We propose a new penalization-based method, called Gava, to tackle this kind of signal by making use of the graphical structure among predictors. The proposed Gava covers several existing methods as special cases. Our numerical examples and theoretical studies demonstrate the effectiveness of the proposed Gava for estimation and prediction.

Session 23: Use of Real-World Data and Real-World Evidence in the Regulatory Decision-making of Medical Products

Collaborative Effort from to use RWD Establish Performance Criteria for medical devices.
Roseann White1, Yu-Ching Cheng2, Tianyi Sun3, Niveditta Rackahara4 and Daniel Bertges5
1Your Third Opinion
2U.S. Food and Drug Administration
3Weill School of Medicine, Cornell University
4Geisel School of Medicine, Dartmouth University
5University of Vermont Medical Center
Roseann.white@your3rdopinion.com

The FDA is building the National Evaluation System for health Technology (NEST) to more efficiently generate better evidence for medical device evaluation and regulatory decision-making. RAPID (The Registry Assessment of Peripheral Interventional Devices) is a NEST demonstration project that was funded through the Medical Device Epidemiology Network (MDEpiNET) a public-private partnership supported by U.S. FDA. The project developed a core minimum set of data elements related to the care and treatment of patients with peripheral arterial disease (PAD). Those elements were extracted from the Vascular Quality Initiative (VQI) registry for procedures performed in 2010 - 2015. VQI is a collaboration of the Society for Vascular Surgery Patient Safety Organization (SVS PSO), 18 regional quality improvement groups organized under the SVS PSO, and Medstreaming/M2S, it’s commercial technology partner to collect details about vascular procedures and outcomes. That data was used to develop performance goals for interventional devices that are used treating patient with PAD. The process that was followed was in line with best practices for statistical analysis of observational data. Prior to looking at the data, the clinical objectives were agreed upon and the SAP was written and signed off before access to the data. Only FDA received a copy of the data from VQI.
Studies
Incorporating Real-World Evidence in Single-Arm Clinical Studies. We present the approach and provide results of simulation studies. We develop the paradigm of making a comparative claim under certain conditions and assumptions. We will briefly review this scenario in the talk. Sometimes, the subject-level data of the RWD source are unavailable, but the statistical models on the relationship between clinical outcomes and covariates have been established based on a large RWD database. In this case, we contend that it is still possible to make a comparative claim if bias issues can be satisfactorily addressed. We will present the approach and provide results of simulation studies.

Making a comparative claim when leveraging information of the control formed from real-world data

Leveraging real-world data (RWD) in medical device development has become popular in recent years. Oftentimes, the clinical performance of the control treatment is captured in the RWD. One common application in the pre-market setting is to conduct a nonrandomized study in which the results of a single-arm investigational study are compared with results of control formed from RWD. It is possible to make a comparative claim if bias issues can be satisfactorily addressed. We will present the approach and provide results of simulation studies.

Incorporating Real-World Evidence in Single-Arm Clinical Studies

Chenguang Wang, Heng Li, Wei-Chen Chen, Nelson Lu, Ram Tiwari, Yunling Xu and Lilly Yue
1 Johns Hopkins University
2 U.S. Food and Drug Administration, CDRH.

cw68@jhmi.edu

We are now at an amazing time for medical product development in drugs, biological products and medical devices. As a result of dramatic recent advances in biomedical science, information technology and engineering, “big data” from health care in the real-world have become available. Although big data may not necessarily be attuned to provide preponderance of evidence to a clinical study, high-quality real world data can be transformed into scientific evidence for regulatory and healthcare decision-making using proven analytical methods and techniques, such as propensity score methodology and Bayesian inference. In this paper, we extend the Bayesian power prior approach for a single-arm study (the current study) to leverage external real-world data. We use propensity score methodology to pre-select a subset of real world data containing patients that are similar to those in the current study in terms of covariates, and to stratify the selected patients together with those in the current study into more homogeneous strata. The power prior approach is then applied in each stratum to obtain stratum-specific posterior distributions, which are combined to complete the Bayesian inference for the parameters of interest. We evaluate the performance of the proposed method as compared to that of the ordinary power prior approach by simulation and illustrate its implementation using a hypothetical example.

Synthesizing Real World Data into Evidence for Medical Devices

Gregory Campbell
GCStat Consulting, LLC
gcstat@verizon.net

In this era of “Big Data” there are many sources of possible data that could be leveraged in the regulatory environment. The effort to synthesize such data into evidence and then utilize such evidence in an efficient manner can be a challenge but the payoff can be enormous in terms of savings of time and resources. There are inherent biases (especially selection biases) of observational data that need to be addressed as well as the confounders to be adjusted for due to different population characteristics. Also at issue is the generalizability of the studies. The potential uses of such real-world evidence (RWE) in the premarket regulatory environment are for expansion of an indication or to serve as data for a control group. The conversion of data into real world evidence can rely on statistical methodologies such as propensity scores and Bayesian statistics. A very real concern in any real-world application is the assessment of the quality of the data. Real-world evidence can also be useful in the post-market in mandated surveillance studies and in condition-of-approval studies. The use of RWE for medical device in the post-market is being leveraged through National Evaluation System for health Technology (NEST), which is coordinated by the Medical Device Innovation Consortium (MDIC), a public-private partnership.

Session 25: Statistical and General Considerations on Development of Gene Therapies

Statistical challenges in the development of gene therapies for rare genetic diseases

Weiliang Shi
Bluebird Bio, Inc.
wshi@bluebirdbio.com

Gene therapies attempt to treat diseases at the molecular level by correcting defective genes or adding functioning genes. Tremendous progress has been made in gene therapies for the treatment of rare genetic diseases. However, there are several major statistical challenges in the development of these therapies. Firstly, randomization may not always be feasible due to the limited patient population, complexity of the treatment process, and ethical considerations. As a result, pivotal studies may have to be single arm, which poses the second challenge - the quantification of effect size associated with standard of care. Historical data in these populations are usually very limited and sponsors may need to conduct their own natural history studies to better understand the use of standard of care. Lastly, the selection of the primary endpoint also needs careful evaluation and consideration. There is typically no precedence in these disease populations and novel endpoints may need to be invented. In this presentation, we will provide an overview of stem cell-based gene therapies, followed by a case study on how these statistical challenges are being addressed.
Regulatory Experiences in the Development and Approval of Gene Therapies
Shiowjen Lee
U.S. Food and Drug Administration, CBER.
Shiowjen.Lee@fda.hhs.gov
Product development and approval for rare diseases have been challenging in many aspects including small patient population size and assessment tools (or endpoints) for clinical benefits. Though there are drugs currently in the market for treating some rare diseases, they at the best are for controlling temporarily or/and delaying symptoms. Gene therapy has been thought a way to fix a genetic problem at its source. The identification of genes responsible for many rare diseases have prompted high hope and belief that gene therapies with gene editing, replacement or delivery of a new gene can effectively treat and even cure diseases. These motivations result in a rapid development of gene therapies for various diseases. In this presentation, I will share regulatory experiences and statistical considerations in the development and approval of gene therapies as well as guidance in development. Several examples will be presented to illustrate our experiences and considerations.

Estimand and patient journey - Analysis considerations for CAR-T Phase 3 studies
Yiyun Zhang
Novartis
yiyun.zhang@novartis.com
The diversity of patient journeys can raise fundamental questions on clinical trial interpretation. The ICH E9 Addendum introduces the concept of an estimand to precisely describe the treatment effect of interest. It provides a structured approach to discuss how to account for intercurrent events that occur after randomization and may affect the assessment or interpretation of the treatment effect. The estimand framework was applied to design a trial for a chimeric anti-gen receptor T cell therapy (CAR-T), which is a novel therapy that involves personalized manufacturing before patients can receive the final product. Unlike traditional drugs, CAR-T requires only one-time infusion and may provide long term benefit. The treatment effect of interest was carefully defined considering the range of patient journeys expected for the study indication and investigational treatment.

Design of Clinical Study for Rare Diseases via Borrowing Historical Data
×Charlie Cao1, Peng Sun1, Ming-Hui Chen2, Yiwei Zhang3, John Zhong1, Guochen Song3 and Zhenxun Wang
1Biogen, Inc.
2University of Connecticut
charlie.cao@biogen.com
In this presentation, a Bayesian design that enables conditional borrowing of historical data will be proposed in a randomized controlled trial setting for the treatment of a rare disease. Conditional borrowing means that borrowing of historical data using power prior occurs only if the difference in sample means between the concurrent control and the historical control falls within a pre-specified range. The key features of the proposed Bayesian design are that, as the distribution of the concurrent control shifts away from that of the historical control, the resulting inflation of Type I error is bounded and the Bayesian estimator of treatment difference remains unbiased. Details of the operational characteristics of the proposed design will be presented and advantage of the proposed design over the Bayesian hierarchical modelling approach will be discussed.

Restricted function-on-function regression models
×Xin Qi and Ruiyan Luo
Georgia State University
xqi@gsu.edu
In historical function-on-function regression model, it is assumed that the value y(t) of the response curve at time t only depends on the values of the predictor curves before this time point, x(s) with s<t. Equivalently, the model can be expressed in terms of the integral of the multiplication of coefficient kernel beta(s,t) and the predictor curve x(s) in a triangular subregion s<t. We consider more general and flexible models, where the integral of the multiplication of coefficient kernel and the predictor curve can be taken in any subregion. Given a subregion, we propose a decomposition of the coefficient kernel and algorithms to jointly estimate the decomposition. When the subregion is unknown, we propose a method to simultaneously determine the subregion and estimate the coefficient kernel.

Functional regression for highly densely observed data with novel regularization
×Ruiyan Luo and Xin Qi
Georgia State University
rliuo@gsu.edu
Modern techniques generate highly densely observed functional data that exhibit complex local variation patterns. To build the scalar-on-function or function-on-function linear regression model for such data, we consider general coefficient functions that can be smooth, nonsmooth, or even discontinuous. The usual smoothness measures, such as the integral of squared derivatives, may not be suitable to differentiate the roughness of these functions. We propose a family of new roughness measures based on the moduli of continuity and wavelet transformation for these general functions. Using these new roughness measures, we propose new regularization and estimation methods for the scalar-on-function and function-on-function linear regression models with highly densely observed functional data. Simulation studies and real data applications illustrate that the new methods have good performance for various coefficient functions and predictor curves. Compared to the smoothness regularization and sparsity regularization, the new regularization is particularly efficient for highly densely spiky functional data. The proposed new regression methods have been implemented in the R package FRegSigCom.

Estimation and Inference for Generalized Geoadditive Models
×Shan Yu1, Guannan Wang2, Li Wang1, Chenhui Liu3 and Lijian Yang4
1Iowa State University
2College of William & Mary
3Turner-Fairbank Highway Research Center
4Tsinghua University
syu@iastate.edu
In many application areas, data are collected on a count or binary response with spatial covariate information. We introduce a new class of generalized geoadditive models (GGAMs) for spatial data distributed over complex domains. Through a link function, the proposed GGAM assumes that the mean of the discrete response variable depends on additive univariate functions of explanatory variables and a bivariate function to adjust for the spatial effect. We propose a two-stage approach for estimating and making inferences of the components in the GGAM. In the first stage, the univariate

Abstracts
components and the geographical component in the model are approximated via univariate polynomial splines and bivariate penalized splines over triangulation, respectively. In the second stage, local polynomial smoothing is applied to the cleaned univariate data to average out the variation of the first-stage estimators. We investigate the consistency of the proposed estimators and the asymptotic normality of the univariate components. We also establish the simultaneous confidence band for each of the univariate components. The performance of the proposed method is evaluated by two simulation studies. We apply the proposed method to analyze the crash counts data in the Tampa-St. Petersburg urbanized area in Florida.

**Distance-Based Analysis with Quantile Regression Models**

Shaoyu Li	extsuperscript{1}, Yangqing Sun	extsuperscript{1}, Liyang Diao	extsuperscript{2} and Xue Wang	extsuperscript{3}

	extsuperscript{1}University of North Carolina at Charlotte

	extsuperscript{2}Seres Therapeutics, Inc.

	extsuperscript{3}Mayo Clinic  

sli23@unc.edu

Suitable pairwise distance measure which defines how dissimilar any two samples are can be used to study the association between multivariate, especially non-vectorially structured multivariate data, which are emerging in many important research areas including genomics, ecology, and neuron imaging. In this work, we consider a quantile regression model for matrices of pairwise distance. We derive large sample properties of estimators in the model and propose a corresponding statistical inference procedure. Intensive simulation studies illustrate great finite sample characteristics of the proposed method in terms of accurate coverage probability, and well controlled empirical type I error rate. Finally, we apply our method to re-analyze a Microbiome association study and a plant ecological study to illustrate its utility.

**Session 27: Statistical Learning for Complex Data**

**High-dimensional Vector Autoregressive Modeling via Tensor Decomposition**

Di Wang	extsuperscript{1}, Heng Lian	extsuperscript{2}, Wai Keung Li	extsuperscript{1} and Guodong Li	extsuperscript{1}

	extsuperscript{1}University of Hong Kong

	extsuperscript{2}City University of Hong Kong  
gdl1@hku.hk

The classical vector autoregressive model is a fundamental tool for multivariate time series analysis. However, it involves too many parameters for high-dimensional time series, and hence suffers from the curse of dimensionality. In this paper, we rearrange the parameter matrices of a vector autoregressive model into a tensor form, and use the tensor decomposition to restrict the parameter space in three directions. Compared with the reduced-rank regression method, which can limit the parameter space in one direction only, the proposed method dramatically improves the capability of vector autoregressive models in handling high-dimensional time series. For this method, its asymptotic properties are studied and an alternating least squares algorithm is suggested. Moreover, for the case with much higher dimension, we further assume the sparsity of three loading matrices, and the regularization method is thus considered for estimation and variable selection. An ADMM-based algorithm is proposed for the regularized method and oracle inequalities for the global minimizer are established.

**A Boosting Algorithm for Estimating Generalized Propensity Scores**

Yeying Zhu	extsuperscript{1}, Donna Coffman	extsuperscript{2} and Debashis Ghosh	extsuperscript{3}

	extsuperscript{1}University of Waterloo

	extsuperscript{2}University of Western Ontario

	extsuperscript{3}University of Waterloo  
yeying.zhu@uwaterloo.ca

We study the causal inference problem with a continuous treatment variable using propensity score-based methods. For a continuous treatment, the generalized propensity score is defined as the conditional density of the treatment-level given covariates (confounders). When the dimension of the covariates is large, the traditional non-parametric density estimation suffers from the curse of dimensionality. Some researchers have suggested a two-step estimation procedure by first modeling the mean function. In this study, we suggest a boosting algorithm to estimate the mean function of the treatment given covariates. In boosting, an important tuning parameter is the number of trees to be generated, which essentially determines the trade-off between bias and variance of the causal estimator. We propose a criterion called average absolute correlation coefficient (AACC) to determine the optimal number of trees. Simulation results show that the proposed approach performs better than a simple linear approximation or L2 boosting. The proposed methodology is also illustrated through the Early Dieting in Girls study, which examines the influence of mothers’ overall weight concern on daughters’ dieting behavior.

**Ensembling Imbalanced-Spatial-Structured Support Vector Machine**

Wenqing He	extsuperscript{1}, Xin Liu	extsuperscript{2} and Grace Yi	extsuperscript{3}

	extsuperscript{1}University of Western Ontario

	extsuperscript{2}Shanghai University of Finance and Economics

	extsuperscript{3}University of Waterloo  
whd@stats.uwo.ca

The Support Vector Machine (SVM) and its extensions have been widely used in various areas due to its great prediction capability. However, research gaps still remain. In particular, these methods cannot effectively handle imbalanced data with a spatial association which commonly arise from many studies such as cancer imaging studies. In this paper, we propose the ensembling imbalanced-spatial-structured support vector machine (EISS-SVM) method which is useful for both balanced and imbalanced data. Not only does the proposed method accommodate the association between the response and the covariates, but also it accounts for the spatial correlation existing in the data. Our EISS-SVM classifier offers a flexible classification tool that embraces the usual SVM as a special case. The proposed method outperforms the competing classifiers, which is demonstrated by simulation studies. The good performance of the proposed method is further confirmed by the application to the real imaging data arising from an ongoing prostate cancer research conducted at the University of Western Ontario, Canada.

**Symmetrization for exact nonparametric functional ANOVA**

Adam B Kashlak

University of Alberta  
Kashlak@ualberta.ca

Testing for equality of means and covariances among functional data groups has received a lot of attention from both parametric approaches via Gaussian processes and nonparametric ones reliant on permutation tests. In this work, we advance nonparametric testing by devising an exact test via a type of Khintchine inequality, a symmetrisation result for random variables in Banach spaces. This approach combines the computational speed of parametric methods with the distribution free benefits of permutation tests. The methodology is very general and can be extended to other data comparisons across categories.
Dynamic correlation analysis and its applications in RNA-seq Data
Tianwei Yu
Emory University

dingj@umn.edu
Dynamic correlations are pervasive in high-throughput data. Large numbers of gene pairs can change their correlation patterns in response to observed/unobserved changes in physiological states. Finding changes in correlation patterns can reveal important regulatory mechanisms. Currently there is no method that can effectively detect global dynamic correlation patterns in a dataset. Given the challenging nature of the problem, the currently available methods use genes as surrogate measurements of physiological states, which cannot faithfully represent true underlying biological signals. In this study we develop a new method that directly identifies strong latent dynamic correlation signals from the data matrix, named DCA: Dynamic Correlation Analysis. At the center of the method is a new metric for the identification of pairs of variables that are highly likely to be dynamically correlated, without knowing the underlying physiological states that govern the dynamic correlation. We validate the performance of the method with extensive simulations. We applied the method to three real datasets: a single cell RNA-seq dataset, a bulk RNA-seq dataset, and a microarray gene expression dataset. In all three datasets, the method reveals novel latent factors with clear biological meaning, bringing new insights into the data.

Whole Genomes Classification and Clustering Based on Nucleotide Positions Distributions
Hsin-Hsiung Huang
University of Central Florida

hsein.huang@ucf.edu
Distributions play an important role for statistical analysis. Natural Vector is an alignment-free method using the position distributions of each nucleotide for whole genome sequences clustering and classification. Recently we investigated the asymptotic theory of natural vectors and proposed a classification method accordingly. Additionally, recently we developed a new representation method of whole genome sequences using Discrete Wavelets Transforms (DWT) and Discrete Wavelets Packet Transform (DWPT), and their asymptotic normality theory. The DWPT Natural Vector has been successfully applied for genomes clustering and classification in several real datasets including very long bacteria whole genome sequences which cannot be handled by other methods such as k-mers, multiple alignments, and Natural Vector. Several examples of clustering and classification of genome sequences have been used for comparisons of these methods.

Estimation of Insulin Resistance using Existing Untargeted Metabolomics Profiling
* Fang-Chi Hsu, Nicholette Palmer, Syh-Huei Chen and Donald Bowden
Wake Forest University

fhsu@wakehealth.edu
Insulin resistance is a central feature of type 2 diabetes (T2D) and overall metabolic health. Insulin resistance is the inability of a person to respond normally to the hormone insulin. While insulin resistance is well-documented as a primary feature of T2D, it has been implicated as a significant contributor to multiple common diseases such as atherosclerosis, Alzheimer’s disease and psychiatric disorders. With few exceptions, the relationship of insulin resistance to disease risk is poorly understood. A major barrier to effectively measuring the contribution of insulin resistance is the difficulty of acquiring high quality measures when the most desirable data requires invasive and expensive examination in a clinical research unit setting. Such data is not commonly acquired, even in diabetes-focused studies, and rarely available in population-based studies of other disease entity. To address this barrier to understanding we will build upon our prior observations that data from contemporary large-scale untargeted metabolomics analysis accounts for a substantial proportion of the inter-individual variation in measures of insulin resistance. With expansive metabolomics becoming increasingly available, this suggests that we can develop estimation methods combining metabolomics and traditional clinical data (e.g. age, sex, BMI) that can accurately estimate measures of insulin resistance in the absence of invasive metabolic testing. We have applied machine learning (e.g. LASSO and Elastic net) methods with 10-fold cross validation to the Insulin Resistance Atherosclerosis Family Study (IRASFS) where insulin resistance and metabolomics have been directly assessed. The correlation between observed and expected insulin resistance ranged from 0.73 to 0.77. We will further validate this model with extension to other cohorts. Thus, effective estimation using metabolomics and clinical measures would provide a means to assess the role of insulin resistance in common diseases.

Session 29: Statistical Machine Learning Methods for Diagnostic Medicine and Biomarkers

A New Information Criterion for Optimal Model Selection
* Jie Ding 1, 2, Yahong Yang 1
1 University of Minnesota
2 Duke University

dingj@umn.edu
We propose 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 miss-specified settings. In practice where the observed data is given without any prior information about the model specification, the proposed criterion is more flexible and reliable compared with classical approaches. We also extend the criterion to high-dimensional variable selection where the sample size is much smaller than the number of variables.

Biomarker models for early Alzheimers disease diagnosis before symptoms appear
Zheyu Wang
Johns Hopkins University

wangzy@jhu.edu

Accumulating evidence suggest that the initiation of the Alzheimer disease (AD) pathogenic process preceed the first symptoms by a decade or more. The recognition of this decade-long asymptomatic stage has greatly impact AD research and therapeutic development to focus on preclinical stage of AD pathogenic process, at which time disease modifying therapy is more likely to be effective. On the other hand, the decade-long preclinical stage imposes 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. Since AD pathophysiology has been recognized as a multidimensional process that involves amyloid deposition, neurotribillary tangles and neurodegeneration among other aspects, we propose latent variable model to study the underlying AD pathophysiology process revealed by multidimensional markers and apply the model to two different AD data sets.

Methods for predicting ovarian cancer diagnosis with a longitudinal biomarker CA-125
Yongli Han1, Ming Wang2 and Danping Liu1
1National Cancer Institute
2Pennsylvania State University
danping.liu@nih.gov

Longitudinal biomarkers are commonly collected for predicting disease outcomes, as the trajectory of a marker carries rich information about the disease risk. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial conducted annual screening of CA-125 on 39,105 women participants, with the interest in predicting ovarian cancer (OC) diagnosis. The risk of ovarian cancer algorithm (ROCA) has been developed in the literature, which models the CA125 marker trajectories separately for the cases and controls, and then derives the risk estimation using Bayes rules. A disadvantage of ROCA is that, it does not make full use of the time of OC diagnosis but dichotomizes the cancer outcome. We propose several conditional models of the longitudinal marker given the time-to-event outcome, which can be estimated from the OC cases only under the assumption of non-informative censoring. The conditional models use flexible kernel functions to characterize the rapid increase of CA-125 shortly before OC diagnosis. Together with a Cox regression of the time-to-event outcome, we calculate the 1-year absolute risk of OC diagnosis, given the longitudinal marker. We demonstrated that the conditional models closely approximate shared random effects models, and is computationally easier to implement. Simulation studies are conducted to examine the performance of the proposed risk calculator in comparison with ROCA, as well as its robustness to model mis-specifications.

Session 30: New Frontier in High-dimensional Data Analysis

Nonregular and Minimax Estimation of Individualized Thresholds in High Dimension with Binary Response
Huijie Feng1, Yang Ning1 and Jiwei Zhao2
1Cornell University
2State University of New York at Buffalo
yn265@cornell.edu

Given a large number of covariates Z, we consider the estimation of a high-dimensional parameter \( \theta \) in an individualized linear threshold \( \theta^T Z \) for a continuous variable \( X \), which minimizes the disagreement between \( \text{sign}(X - \theta^T Z) \) and a binary variable \( Y \). While the problem can be casted into the M-estimation framework, minimizing the corresponding empirical risk function is computationally intractable due to nonsmoothness of the loss function. Moreover, estimating \( \theta \) even in the fixed dimensional setting is known as a nonregular problem leading to nonstandard asymptotic theory. To tackle the computational and theoretical challenges in the estimation of the high-dimensional parameter \( \theta \), we propose an empirical risk minimization approach based on a regularized smoothed loss function. To guarantee the Fisher consistency of the proposed method, the surrogate smoothed loss function has to be non convex in general. A computationally efficient path-following algorithm is proposed to compute the estimator. The computational and statistical trade-off is studied. Statistically, we show that the finite sample error bound for estimating \( \theta \) in \( l_2 \) norm is \( (s \log d/n)^{\beta/(2\beta + 1)} \), where \( s \) is the sparsity of \( \theta \), \( n \) is the sample size and \( \beta \) is the smoothness of the conditional density of \( X \) given the covariates \( Z \). The convergence rate is nonstandard and slower than that in the classical Lasso problems. Furthermore, we prove that the resulting estimator is minimax rate optimal up to a logarithmic factor. The Lepski method is developed to achieve the adaption to the unknown sparsity \( s \) and smoothness \( \beta \). Computationally, we show that the path-following algorithm achieves geometric rate of convergence for computing the whole regularization path. Finally, we evaluate the finite sample performance of the proposed estimator in simulation studies and a real data analysis from the ChAMP (Chondral Lesions And Meniscus Procedures) Trial.

High-Dimensional Robust Thresholded Regression with Application to Scalar-on-Image Analysis
Lingzhou Xue
Pennsylvania State University
lxue@psu.edu

In this talk, we introduce a new class of high-dimensional robust thresholded regression for estimating sparse linear models with a complex dependence structure in predictors (such as the imaging data). We propose an efficient nonconvex estimation procedure based on thresholding functions and robust Huber loss. The proposed method provides a powerful tool to analyze large-scale data analysis.
sets with complex dependence and possible outliers. After analyzing the landscape of the proposed method, we establish both statistical and computational consistency without requiring the restrictive incoherence condition under the high-dimensional setting. The numerical properties are demonstrated in simulation studies. Furthermore, we present a real application to scalar-on-image analysis using resting-state functional magnetic resonance imaging data from the Autism Brain Imaging Data Exchange study.

**BET on Independence**  
**Kai Zhang**  
University of North Carolina at Chapel Hill  
[zhangk@email.unc.edu](mailto:zhangk@email.unc.edu)

We study the problem of nonparametric dependence detection. Many existing methods may 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 symmetry statistics in the filtration are complete sufficient statistics for dependence. These statistics are also uncorrelated under the null. By utilizing symmetry statistics, 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 symmetry 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.

**OPTIMAL ADAPTIVITY OF SIGNED-POLYGON STATISTICS FOR NETWORK TESTING**  
**Tracy Ke**  
Harvard University  
[tracy@fas.harvard.edu](mailto:tracy@fas.harvard.edu)

Given a symmetric social network, we are interested in testing whether it has only one community or multiple communities. The desired tests should (a) have a tractable limiting null distribution, (b) accommodate possible severe degree heterogeneity and mixed-memberships, (c) adapt automatically to different levels of sparsity, and (d) achieve the optimal phase diagram. How to find such a test is a challenging problem. We propose the Signed Polygon as a class of new tests. Fix $m$ 3. For each m-gon in the network, we define a score using the centralized adjacency matrix, and the m-th order Signed Ploygon statistic is the sum of such scores. The Signed Polygon include the Signed Triangle (SgnT) and Signed Quadrilateral (SgnQ) as special examples. We show that both the SgnT and SgnQ tests satisfy (a)-(d), and especially, they work well for both the very sparse and less sparse cases.

**Trends in extreme value indices**  
**Laurens Dehaan and Nan Zou**  
Erasmus University Rotterdam  
[zhou@ese.eur.nl](mailto:zhou@ese.eur.nl)

We consider extreme value analysis for independent but non-identically distributed observations. In particular, the observations do not share the same extreme value index. Assuming continuously changing extreme value indices, we provide a non-parametric estimate for the functional extreme value index. Besides estimating the extreme value index locally, we also provide a global estimator for the trend and its joint asymptotic theory. The asymptotic theory for the global estimator can be used for testing a pre-specified parametric trend in the extreme value indices. In particular, it can be applied to test whether the extreme value index remains at a constant level across all observations.

**Session 31: Extreme value statistics**

Towards inference for multivariate time series extremes: multiple block sizes and overlapping blocks  
**Axel Bucher$^1$, Stanislav Volgushev$^2$ and Nan Zou$^2$**  
$^1$Heinrich-Heine-Universität Düsseldorf  
$^2$University of Toronto  
[nan.zou@utoronto.ca](mailto:nan.zou@utoronto.ca)

The classic block maxima method aims to estimate the extreme value distribution with maxima over blocks of data. The errors in this estimation are two-fold. First, the extreme value distribution of interest characterizes the maxima over an infinitely long block. However, given finite amount of data, one can at most get the distribution of the maxima over a finitely long block. This difference between infinite and finite block length causes a deterministic bias. Second, the distribution of the maxima over a finitely long block is deterministic. Given the randomness of data, one can at most come up with an estimation. This difference between an estimation and the truth gives a stochastic error. In a generic multivariate weakly dependent setting, a new estimator is proposed to diminish these two errors. To bring down stochastic error, we apply sliding blocks and take an average over multiple block sizes. To reduce the deterministic error, we estimate and remove the dominating term of bias. The resulting estimator has been proven to enjoy consistency, asymptotic normality, a variance of a smaller size, and a bias of a smaller order.

**Dynamic Multivariate Peak over Threshold Model**  
**Zifeng Zhao**  
University of Notre Dame  
[zzhao@nd.edu](mailto:zzhao@nd.edu)

In this paper, we propose a dynamic multivariate Peak over Threshold (dmPoT) framework for modeling the time-varying behavior of joint tail risks among multiple financial time series. The dmPoT model generalizes the popular univariate dynamic PoT approach to the multivariate setting, and thus provides a direct and simultaneous modeling of the time-varying marginal and joint behavior of tail risks. The proposed model offers a new angle to study the tail dependence dynamics in financial markets. Extensive numerical experiments demonstrate the flexibility of dmPoT model and confirm the reliability of its maximum likelihood estimation. Two real data examples demonstrate the promising applications of dmPoT on revealing the tail dynamics of global financial markets and on portfolio optimization for S&P sector indices.

**Session 32: Longitudinal Data Analysis with Finite Mixtures**

On the improved estimation of the normal mixture components for longitudinal data  
**Tapio Nummi**  
Tampere University  
[Tapio.Nummi@tuni.fi](mailto:Tapio.Nummi@tuni.fi)
In many applications of data modeling a finite normal mixture (see e.g. Everitt and Hand, 1981) as well as its extensions, such as skew-normal distribution (Lin et al. 2007b) or t-distribution (Lin et al. 2007a), provides a sensible model for the data at hand. However, finding the best possible component distribution among many competing alternatives can be a very complicated and challenging task. For example, the number of components needed may depend on the sample size, assumed component distribution as well as the possible transformation (scale of measurements) applied. Actually, two slightly different goals can be identified: Approximation of the distribution of the response variable and identification of the number of sub-populations. The focus of this talk is on trajectory analysis (TA) that is an application of finite mixture modeling for longitudinal data (Nagin, 1999 and 2005). We propose a method that is based on the scaled Box-Cox transformation (Spitzer, 1984 and Gurka et al. 2006) that makes the likelihood based inference possible over transformed responses and the actual analyses are then based on the components of transformed normal mixtures. The proposed approach has several advantages. First, the theory of TA with normal mixtures is well established and a number of implementations as software packages already exists. Second, a suitable transformation can reduce the risk of generating the excess number of trajectory groups. The data analytic part of the talk is based on real data mixture applications of birth weight and trajectories of alcohol consumption. In the later application, a zero inflation problem with continuous longitudinal measurements is also discussed. References: Everitt, B. and Hand, D. (1981). Finite Mixture Distributions. London: Chapman and Hall. Gurka, M., Edwards, L., Muller, K. and Kupper, L. (2006). Extending the Box-Cox transformation to the linear mixed model. Journal of the Royal Statistical Society, A (Statistics in Society), Volume 169, Issue 2, p. 273-288. Lin, T., Lee, J. and Hsieh, W. J. (2007a). Robust mixture modeling using the skew t-distribution. Statist. Comput., 17, p. 81-92. Lin, T., Lee, J. and Yen, S. (2007b). Finite mixture modeling using the skew normal distribution. Statistica Sinica, 17, p. 909-927. Nagin, D. (1999). Analyzing developmental trajectories: Semi-parametric, group-based approach. Psychological Methods, 4, p. 139-177. Nagin, D. (2005). Group-based modeling of development. Cambridge, MA: Harvard University Press. Spitzer, J. (1984). Variance estimates in models with the Box-Cox transformation: Implications for estimation and hypothesis testing. Review of Economics and Statistics, 66, p. 645-652.

Survival analysis with finite mixtures

Janne Salonen and Jyrki Miettinen

1 Finnish Centre for Pensions
2 University of Helsinki

janek.salonen@gmail.com

The survival analysis techniques are widely used in many fields of science. Survival analysis is capable of censoring, that is, handling incomplete information. Observations are censored when the information about their survival time is incomplete. Models based on a hazard function have dominated survival analysis since Cox constructed the proportional hazard model. Frailty models are a specific area in survival analysis. Frailty provides a way to introduce unobserved heterogeneity into models for survival data. A univariate frailty model handles heterogeneity in survival analysis. The key idea is that individuals have different frailties. A shared frailty model provides an extension of the univariate frailty model, and it allows a mutual dependence of clustered mortalities to be taken into account in the analysis. The finite mixture analysis aims to understand the heterogeneity in the population. One source of heterogeneity may arise from possible sub-populations. Finite mixture modelling tools are used in many fields of science but, so far, they are not that well known in survival analysis. We intend to combine survival analysis with new tools of finite mixture analysis. Using pension insurance mortality data, we seek to model mixture components under the Cox survival regression model.

Practical Modelling and Design Aspects of Nonlinear Trajectory Analysis

Timothy Obrien and Holly Michalak

Loyola University Chicago

teobrien@gmail.com

Longitudinal data are ubiquitous in biomedical research, economics, environmental research, psychometrics as well as many other domains, and analysis of these data present unique and far-reaching challenges in applied statistical research. These data often also contain latent (hidden) cohorts/groups, which - with the aid of the EM algorithm and associated methods - can be discerned in order to help researchers in better understanding their data and underlying phenomena. Although the fields of Finite Mixture Models and Trajectory Analysis in the context of longitudinal data analysis is relatively new, controversy exists as to how best to discern these patterns and data. This talk focuses on trajectory analysis and finite mixture models in modelling nonlinear phenomena and focusing on estimation and design. We make connections to the linear and generalized linear cases - as well as highlighting important differences and relevant software packages.

Session 33: Recent advances in modern survival analysis and the novel applications

Multiplicative Rates Model for Recurrent Events in Case-Cohort Studies

Jianwen Cai 1, Poulami Maitra 1 and Leila Daf Amorim 2

1 University of North Carolina at Chapel Hill
2 Federal University of Bahia

cal@bios.unc.edu

In large prospective cohort studies, accumulation of covariate information and follow-up data make up the majority of the cost involved in the study. This might lead to the study being infeasible when there are some expensive variables and/or the event is rare. Prentice (1986) proposed the case-cohort study for time to event data to tackle this problem. There has been extensive research on the analysis of univariate and clustered failure time data, where the clusters are formed among different individuals under case-cohort sampling scheme. However, recurrent event data are quite common in biomedical and public health research. We propose case-cohort sampling schemes for recurrent events. We consider multiplicative rate models for the recurrent events and propose weighted estimating equations approach for parameter estimation. We showed that the estimators are consistent and asymptotically normally distributed. The proposed estimator performed well in finite samples in our simulation studies. For illustration purposes, we examined the association between prior occurrence of measles on Acute Lower Respiratory Infection (ALRI) among young children in Brazil.

Personalized Treatment Selection for Joint Optimization of Survival and Other Outcomes

Sommath Datta

University of Florida

sommath.datta@ufl.edu

In this work, we propose a novel method for individualized treatment selection when the treatment response is multivariate which
includes a survival component that is subject to right censoring. Since our method covers arbitrary number of treatments and outcome variables, it can be applied to a broad set of models. In addition, the performance measure for each component response can be adjusted depending on the nature of the response. As for example, for a survival component, we might use the difference of mean survivals whereas, for some other clinical covariate, a difference of means may be more suitable. The proposed joint optimization method uses a rank aggregation technique to estimate an ordering of treatments based on ranked lists of treatment performance measures. The method has the flexibility to incorporate patient and clinician preferences to the optimal treatment decision on an individual case basis. An empirical study demonstrates the performance of the proposed method in finite samples. We also present a data analysis using a HIV clinical trial data to show the applicability of the proposed procedure for real data.

Simultaneous Estimation and Variable Selection for Interval-censored Failure Time data
*(Tony) Jianguo Sun*
University of Missouri
SunJ@Missouri.edu

Simultaneous Estimation and Variable Selection for Interval-censored Failure Time data Abstract: In this talk, we will discuss regression analysis of interval-censored failure time data under the Cox model with the focus on simultaneous estimation and variable selection. For the problem, a new method, Broken Adaptive Ridge Regression, will be presented and both finite sample and asymptotic studies are carried out to evaluate the performance of the presented approach.

Session 34: Statistical Issues on Developing Medicines for Rare Diseases or Vulnerable Populations

The Use of Historical Data in Assessing the Efficacy of Drugs in Rare Diseases
*Ros Walley*
UCB Pharma, Inc.
rwalley@ucb.com

The newly revised FDA guidance: “Rare Diseases: Common Issues in Drug Development” restates the usual requirement to demonstrate substantial evidence of a drug’s effectiveness and provide sufficient safety information for approval of any drug, regardless of the prevalence of the disease. It then goes on to describe the flexibility the FDA regulations provide in applying these standards and includes a section on the use of historical controls. The EMA/CHMP Guideline on Clinical Trials in Small Populations is that “there are no special methods for designing, carrying out or analysing clinical trials in small populations. There are, however approaches to increase the efficiency of clinical trials. The need for statistical efficiency should be weighed against the need for clinically relevant / interpretable results; the latter being the most important”. In this talk we investigate the role of historical data in increasing the efficiency of clinical trials in rare diseases and paediatric populations. In the case of the rare diseases this may be restricted to the use of historical control data and we review Bayesian approaches to incorporate such data formally in the design and analysis of studies. For paediatric populations, in addition to the use of historical control data, it may be possible to extrapolate from an adult population to children. In all cases the need to assess potential prior-data conflicts is of paramount importance. In this talk we review the current status of these approaches. References: FDA guidance: Rare Diseases: Common Issues in Drug Development Guidance for Industry. https://www.fda.gov/drugs/guidancemain/industryinformation/guidancedocuments/ucm423881.htm European Medicines Agency (2006). Guideline on clinical trials in small populations. https://www.ema.europa.eu/en/clinical-trials-small-populations

Design of a Phase 3 Trial for an Acute Treatment of a Rare Disease with Episodic Attacks
*Sharon Murray*
BioCryst Pharmaceuticals, Inc.
smurray@biocryst.com

There are many clinical, regulatory, and statistical considerations that factor into the design of a Phase 3 trial. Should a crossover or parallel trial be used? What endpoints should be considered? Which should be primary and which should be secondary? For treatment of a rare disease, there is also the consideration of minimizing the number of subjects while meeting the regulatory requirements. Crossover trials can reduce the number of subjects required while parallel trials eliminate the chance of any carryover effect. Of course, feedback from regulatory agencies such as the FDA must be accounted for. Design considerations will be discussed using as an example the design of a Phase 3 trial for an acute treatment of a rare disease with episodic attacks. The benefits of crossover vs. parallel design in this setting will be discussed along with sample size and endpoint considerations. Highlights from the FDA draft guidance on rare diseases will be provided.

Challenges, Achievements and Lessons Learned from two Clinical Studies in Women of Child Bearing Age
*Maggie Wang*
UCB Biosciences, Inc.
maggie.wang@ucb.com

Women of reproductive age with chronic inflammatory diseases face much uncertainty regarding the safety of using biologics. However, there are very limited clinical study data available targeting these women due to complexities in addition to legal and ethical challenges. Studying how biologic drugs impact these women can bring more choices to this vulnerable population with more treatment options. Two key aspects of these clinical studies are studying the transfer of a biologic drug through the placenta to the fetus and any potential drug transfer via the mother’s breast milk to the infant. Anti-TNF lactation data historically has been based on case reports without any systematically designed studies. We conducted systematically controlled clinical trials to measure the presence of an anti-TNF in breastfeeding and infants. The goal is to provide scientifically validated data to enable mothers making informed decisions and improving their quality of life during pregnancy and breastfeeding. Study design, operational considerations and challenges for two clinical trials will be presented and discussed.

Session 35: New frontiers in change point and complex data analyses

Empirical Likelihood for Change Point Detection in Autoregressive Models
*Ramadha Piyadigamage* and ♦Wei Ning
1Western Washington University
2Bowling Green State University

Change point analysis has become an important research topic in many fields of applications. Several research work have been car-
ried out to detect changes and its locations in time series data. In this paper, a nonparametric method based on the empirical likelihood is proposed to detect structural changes in the parameters of autoregressive (AR) models. Under certain conditions, the asymptotic null distribution of the empirical likelihood ratio test statistic is proved to be Gumbel type. Further, the consistency of the test statistic is verified. Simulations are carried out to show that the power of the proposed test statistic is significant. The proposed method is applied to monthly average soybean sales data to further illustrate the testing procedure.

**Robust Graph Change-point Detection for Brain Evolvement Study**

Fang Han\(^1\), Xi Chen\(^2\), \textbullet{} Honglang Wang\(^3\), Lexin Li\(^4\) and Brain Caffo\(^5\)

\(^1\)University of Washington  
\(^2\)New York University  
\(^3\)Indiana University-Purdue University Indianapolis  
\(^4\)University of California, Berkeley  
\(^5\)Johns Hopkins University

This paper studies brain structural evolvement from resting-state functional magnetic resonance imaging. The brain structure is characterized by a series of Gaussian graphical models, and we propose a robust data-driven method for inferring the structural changes of multiple graphs. The graphs correspond to different subjects, and need to be estimated from the subject level data. We propose to estimate the structural changes of these graphs through a three-step procedure. First, we employ a kernel-smoothing approach to estimate multiple graphs at different ages simultaneously. Secondly, we summarize graphical information, such as the number of edges, global and local efficiency, for each estimated graph, and align them as a curve. Lastly, we propose a robust least-absolute-deviation (LAD) type penalization procedure with the fused Lasso (FL) penalty, named LAD-FL, to infer the change-points in those graph summary metrics. Our method is theoretically well understood, and results show that it could effectively capture the brain evolvement pattern.

**Testing Community Structure for Hypergraphs**

Mingao Yuan\(^1\), Ruiqi Liu\(^1\), Yang Feng\(^2\) and Zuofeng Shang\(^2\)

\(^1\)North Dakota State University  
\(^2\)Indiana University-Purdue University Indianapolis

Many complex networks in the real world can be formulated as hypergraphs where community detection has been widely used. However, the fundamental question of whether communities exist or not in an observed hypergraph still remains unresolved. The aim of the present talk is to tackle this important problem. Specifically, we study when a hypergraph with community structure can be successfully distinguished from its Erdős-Rényi counterpart.

**Design based incomplete U-statistics**

Xiangshun Kong\(^1\) and \textbullet{} Wei Zheng\(^2\)

\(^1\)Beijing Institute of Technology  
\(^2\)University of Tennessee

U-statistics are widely used in the fields of economy, machine learning and statistics. While it enjoys very desirable statistical properties, it also admits an obvious draw back: the computation easily becomes impractical as the data size \(n\) increases. Particularly, the number of combinations, say \(m\), that a U-statistic of order \(d\) has to evaluate is in the order of \(O(n^d)\). Many efforts have been made to approximate the original U-statistic by a small subset of the combinations in history since Blom (1976), who has coined such an approximation as an incomplete U-statistic. To the best of our knowledge, all existing methods require \(m\) to grow at least faster than \(n\), albeit much slower than \(n^d\), in order for the corresponding incomplete U-statistic to be asymptotically efficient in the sense of mean squared error. In this paper, we introduce a new type of incomplete U-statistics, which can be asymptotically efficient even when \(m\) grows slower than \(n\). In some cases, \(m\) is only required to grow faster than \(\sqrt{n}\). Both theoretical and empirical results show significant improvements on the statistical efficiency by the new incomplete U-statistic.

**Session 36: Statistical Inference for Multi-layer data with Complex Dependence**

**Inference on Multi-level Partial Correlations based on Multi-subject Time Series Data**

Yumou Qiu

Iowa State University

yumouqiu@iastate.edu

Partial correlations are commonly used to analyze the conditional dependence among variables. In this work, we propose a hierarchical model to study both the subject and population level partial correlations based on multi-subject time series data. Multiple testing procedures adaptive to temporally dependent data with false discovery proportion control are proposed to identify the nonzero partial correlations in both the subject and population levels. A computationally feasible algorithm is developed. Theoretical results and simulation studies demonstrate the good properties of the proposed procedures. We illustrate the application of the proposed methods in a real example of brain connectivity on fMRI data from normal healthy persons and patients with Parkinson’s disease.

**Hierarchical Functional Data Analysis**

Yuhang Xu

University of Nebraska-Lincoln

yuhangxu86@gmail.com

In plant phenome studies, data are often collected over a period of time and have a hierarchical structure. We propose a spline-based nonparametric approach to perform hierarchical functional principal components analysis for such phenotypic data. We estimate the covariance functions of the functional random effects by a fast penalized tensor product spline approach, perform multi-level functional PCA using the BLUP, and improve the efficiency of mean estimation by iterative decorrelation. We select the number of principal components using a data-driven method. The methods are illustrated with a simulation study and phenotypic data.

**Yield forecasting based on short time series with high spatial resolution data**

Sayli Pokal, \textbullet{} Yuzhen Zhou and Trenton Franz

University of Nebraska-Lincoln

yuzhenzhou@unl.edu

Crop yield forecasting plays an important role in planning and management of fields. Yet, it becomes especially challenging to do forecasting when we only have short time series. The corn yield data in this study were collected in high spatial resolution (i.e., 10 x 10 m spatial resolution in 800 x 800 m domain) every other year from 2002 to 2016. In this paper, we propose a clustering based spatially varying auto-regressive model for forecasting yield. We compare
the forecasting performance of our model with traditional time series model and a few machine learning algorithms. The results show that for short time series with high spatial resolution data, our proposed model outperforms other models.

Session 37: Modern Perspectives and Recent Advances in Quantile Regression

Quantile regression of recurrent events risk
Huitjuan Ma, *Limin Peng, Chiang-Yu Huang and Haoda Fu
1Emory University
limin@emory.edu
Progression of chronic disease is often manifested by repeated occurrences of disease-related events over time. Delineating the heterogeneity in the risk of such recurrent events can provide valuable scientific insight for guiding customized disease management. In this paper, we propose a new sensible measure of individual risk of recurrent events and present a quantile regression framework thereof, which naturally accounts for both observed covariates and unobservable frailty. The proposed modeling requires no distributional specification of the unobservable frailty, while permitting the exploration of dynamic effects of the observed covariates. We develop estimation and inference procedures for the proposed model through a novel adaptation of the principle of conditional score. The asymptotic properties of the proposed estimator, including the uniform consistency and weak convergence, are established. Extensive simulation studies demonstrate satisfactory finite-sample performance of the proposed method. We illustrate the practical utility of the new method via an application to a diabetes clinical trial that explores the risk patterns of hypoglycemia in Type 2 diabetes patients.

Counterfactual Analysis Using Quantile Regression
Yonggang Yao
SAS Institute, Inc.
yonggang.yao@sas.com
Quantile regression can estimate probability distribution of a response variable conditional on the explanatory covariates. To perform quantile-regression counterfactual analysis, you can first build the conditional distribution model by using only the control-group observations, and then predict the conditional distributions for the treatment-group observations by using this model. If all the confounding effects are appropriately included in this model, the predicted distributions are counterfactual that estimate the distributions of the unobservable treatment-group responses as if they were in the control group. This work describes a standardized process for (1) predicting the counterfactual marginal distribution of the treatment-group response, (2) visualizing treatment effect and selection bias by plotting the observed control-group distribution, the observed treatment-group distribution, and the predicted treatment-group counterfactual distribution, (3) comparing the these distributions by using Mann-Whitney-Wilcoxon test. This work also reports the results of the relevant numerical studies.

Model Based Joint Quantile Regression with Dependency
Surya Tokdar
Duke University
surya.tokdar@duke.edu
Quantile regression is widely seen as an ideal tool to understand complex predictor-response relations. Its biggest promise rests in its ability to quantify whether and how predictor effects vary across response quantile levels. But this promise has not been fully met due to a lack of statistical estimation methods that perform a rigorous, joint analysis of all quantile levels. This gap has been recently bridged by Yang and Tokdar (JASA, 2017). Here we present extension of this model to a more comprehensive regression analysis framework that accounts for various kinds of dependence between response observations, including dependent multivariate response and spatiotemporal dependence. We also develop a random effects extension of quantile regression by framing it as a question of modeling within-group dependence.

Distributed Inference for Quantile Regression Processes
Stanislav Volgushev1, *Shih-Kang Chao2 and Guang Cheng3
1University of Toronto
2University of Missouri
3Purdue University
chaosh@missouri.edu
The increased availability of massive data sets provides a unique opportunity to discover subtle patterns in their distributions, but also imposes overwhelming computational challenges. To fully utilize the information contained in big data, we propose a two-step procedure: (i) estimate conditional quantile functions at different levels in a parallel computing environment; (ii) construct a conditional quantile regression process through projection based on these estimated quantile curves. Our general quantile regression framework covers both linear models with fixed or growing dimension and series approximation models. We prove that the proposed procedure does not sacrifice any statistical inferential accuracy provided that the number of distributed computing units and quantile levels are chosen properly. In particular, a sharp upper bound for the latter are derived to capture the minimal computational cost from a statistical perspective. As an important application, the statistical inference on conditional distribution functions is considered. Moreover, we propose computationally efficient approaches to conducting inference in the distributed estimation setting described above. These approaches directly utilize the availability of estimators from subsamples and can be carried out at almost no additional computational cost. Simulations confirm our statistical inferential theory.

Session 38: Object Oriented Data Analysis

Local Hypothesis Testing for Functional Data: Extending FWER and FDR to the Functional Framework
Niels Asken Landtorp Olsen1, Alessia Pini2 and Simone Vantini3
1University of Copenhagen
2Università Cattolica del Sacro Cuore (Italy)
3Politecnico di Milano
simone.vantini@polimi.it
A topic which is becoming more and more popular in Functional Data Analysis is local inference, i.e., the continuous statistical testing of a null hypothesis along the domain. The principal issue in this topic is the infinite number of tested hypotheses, which can be seen as an extreme case of the multiple comparisons problem. During the talk we will define and discuss the notion of Family-wise Error Rate (FWER) and False Discovery Rate (FDR) in the setting of functional data defined on a compact set. We will then introduce two procedures (i.e., the interval-wise testing and a continuous version of the Benjamini-Hochberg procedure) able to control the FWER and the FDR over the functional domain, respectively, and finally describe their inferential properties in terms of control of the Type-I error probability and of consistency. The proposed method will be applied to satellite measurements of Earth temperature with
the aim of identifying the regions of the planet where temperature has significantly increased in the last decades.

**PCA and Asymptotics on the Torus with Application to Biomolecules**

*Benjamin Eltzner1, Stephan Huckemann1 and Kanti Mardia2*

1University of Goettingen
2University of Leeds and University of Oxford [beltzne@uni-goettingen.de](mailto:beltzne@uni-goettingen.de)

Data on a flat torus of arbitrary dimension presents an intricate challenge to statistical data analysis. The cyclic topology of the torus has the consequence that almost all geodesics lie dense in the torus. Therefore, geodesics are not a suitable tool for data analysis and dimension reduction. Local approaches like tangent space PCA struggle with the problem that they have to cut open the torus at some point and can thus lead to dubious data representation, especially if data are very spread out, as is frequently the case for Proteins and RNA. We present some of the challenges associated to data on the torus and a PCA approach which seeks to carefully deal with the problems. We will briefly discuss asymptotics and present results for biomolecular data.

**Analysis of Populations of Networks: Structure Spaces and the Computation of Summary Statistics.**

*Anna Calissano1, Aasa Feragen2 and Simone Vantini3*

1Politecnico di Milano
2University of Copenhagen [anna.calissano@polimi.it](mailto:anna.calissano@polimi.it)

The statistical analysis of network structures is a very challenging topic in terms of thrilling modelling requirements and important applications in many fields such as genomics, proteomics, neuroscience, mobility, and ICT. Despite of this, while the research concerning the analysis of a single network (i.e. network analysis) is a very lively and matured field of investigation, a rigorous statistical modelling of one or more populations of networks (i.e., Object-oriented Data Analysis for network data) is still little developed. In this talk, we are going to describe a statistical modelling of populations of networks based on the notion of Structure Space (Jain and Obermayer, 2009). This space is a finite dimensional quotient space indentifying networks to be the equivalent up to nodes' permutations. In this framework, the permutation action is non trivial, thus the Structure Space is not a manifold. New paradigms are required to compute on Structure Spaces even basic summary statistics such as the Fréchet Mean and Variance, or to perform explorative analysis like Principal Component Analysis. In this talk, we are going to describe a possible approach aimed at soundly perform tasks like the ones just mentioned. The theoretical ideas are going to be clarified by a real world application concerning public transportation system comparison. A set of subway networks in different cities around the world is statistically analysed and the findings discussed.Jain, Brijnesh J., and Klaus Obermayer. “Structure spaces.” Journal of Machine Learning Research 10.Nov (2009): 2667-2714.

**Session 39: Student Paper Award Session**

**Geostatistical Modeling of Positive Definite Matrices Using the Spatial Wishart Process**

*Zhou Lan1, Brian J Reich1, Joseph Guinness2 and Dipankar Bandyopadhyay3*

1North Carolina State University
2Cornell University
3Virginia Commonwealth University [zlan@ncsu.edu](mailto:zlan@ncsu.edu)

Geostatistical modeling for continuous point-referenced data has been extensively applied to neuroimaging because it produces efficient and valid statistical inference. However, diffusion tensor imaging (DTI), a neuroimaging characterizing the brain structure produces a positive definite (p.d.) matrix for each voxel. Current geostatistical modeling has not been extended to p.d. matrices because introducing spatial dependence among positive definite matrices properly is challenging. In this paper, we use the spatial Wishart process, a spatial stochastic process (random field) where each p.d. matrix-variate marginally follows a Wishart distribution, and spatial dependence between random matrices is induced by latent Gaussian processes. This process is valid on an uncountable collection of spatial locations and is almost surely continuous, leading to a reasonable means of modeling spatial dependence. Motivated by a DTI dataset of cocaine users, we propose a spatial matrix-variate regression model based on the spatial Wishart process. A problematic issue is that the spatial Wishart process has no closed-form density function. Hence, we propose approximation methods to obtain a feasible working model. A local likelihood approximation method is also applied to achieve fast computation. The simulation studies and real data analysis demonstrate that the working model produces reliable inference and improved performance compared to other methods.

**Integrative Analysis of Irregularly Measured and Mixed-Type Biomarkers in Electronic Health Records**

*Jitong Lou1, Yuanjia Wang2, Lang Li3 and Donglin Zeng1*

1University of North Carolina at Chapel Hill
2Columbia University
3Ohio State University [jitong@live.unc.edu](mailto:jitong@live.unc.edu)

Electronic health records (EHRs) has increasingly become an important data source for personalized medicine. In EHRs, disease biomarkers from the same patient are recorded longitudinally at clinical encounters. In order to comprehensively assess patient’s disease comorbidity and susceptibility, it is necessary to characterize these biomarkers over time in an integrative way. However, there exist some challenges such as, in EHRs, the biomarkers are measured sparsely at irregular and informative clinical encounters. In this paper, we propose multivariate generalized linear models to analyze mixed-type biomarkers, where latent multivariate Gaussian temporal processes are introduced to capture between-marker dependence over time. We allow the covariate effects and covariance matrix of the latent processes to be time-dependent. For inference, we apply kernel-weighted estimating equations based on the method of moments where kernel weights account for the heterogeneous intensity of measurement times. We investigate the asymptotic properties of the derived estimators. Under the multivariate models, we integrate the irregularly measured biomarkers of mixed modes into composite scores that reflect patients’ underlying health status. We illustrate the finite-sample performance of our method through extensive simulation studies. Lastly, we apply our method to analyze a large sample of EHRs of Type 2 Diabetes (T2D) patients and show its utility to characterize patients’ comorbidity and disease progression while accounting for challenges of EHRs.Keywords: Electronic health records, latent process, kernel smoothing, generalized linear models, method of moments, type 2 diabetes

**GWAS-based AMD Progression Using a Copula Semiparametric Model**

*Tao Sun, Wei Chen and Ying Ding*
University of Pittsburgh

tao.sun@pitt.edu

The genome-wide association studies (GWAS) of Age-related Macular Degeneration (AMD), a progressive bilateral eye disease, is the first and most successful GWAS research, where the massive GWAS data provide unprecedented opportunities to study disease risk and progression. This research is motivated by discovering genetic causes and making accurate prediction for AMD progression. For genetic variant identification, we develop a copula-based semiparametric approach for modeling and testing bivariate censored data. Specifically, the joint likelihood is modeled through a two-parameter Archimedean copula, which can flexibly characterize the dependence between two margins. The marginal distributions are modeled through a semiparametric transformation model using sieves, with the proportional hazards or odds model being a special case. We propose a sieve maximum likelihood estimation procedure and develop a generalized score test for testing the regression parameter(s). We apply our method to a genome-wide analysis of AMD progression to identify susceptible risk variants for the disease progression. Lastly, we build a novel GWAS-based survival neural network prediction model for AMD progression. Our results demonstrate how the synergy of wealthy GWAS data and deep learning can effectively predict survival probabilities.

Semiparametric Modelling and Estimation of the Global Per-centile Outcome

*Xiangyu Liu*, Jing Ning, Xuming He and Ruoshua Li
1University of Texas Health Science Center at Houston
2University of Texas MD Anderson Cancer Center
3University of Michigan

Biomedical studies are often challenged by the lack of a single primary outcome that can comprehensively capture the multidimensional impairments and symptoms of a disease. To address this, global composite outcomes are commonly formulated to integrate multiple individual outcomes and to achieve comprehensive assessments of the overall disease severity and burden. The global rank-sum outcome and the corresponding K-sample test procedures have been successfully applied in many clinical studies, but existing methods do not lend themselves to regression modeling. In this work, we consider sensible regression strategies to evaluate covariate effects on the global percentile outcome (GPO), under the transformed linear model and the monotonic index model respectively. Posing minimal and realistic assumptions, we develop estimation and inference procedures that account for the special features of the GPO. Asymptotic results were established rigorously using U-statistic and U-process techniques. Simulation studies suggest that the proposed methods perform satisfactorily under realistic sample sizes. An application to a Parkinson’s disease dataset illustrates the practical utilities of the proposed methods.

**BART with Targeted Smoothing: an Analysis of Patient-Specific Stillbirth Risk**

Jennifer Starling
University of Texas at Austin

We introduce BART with Targeted Smoothing, or tsBART, a new Bayesian tree-based model for nonparametric regression. The goal of tsBART is to introduce smoothness over a single target covariate \(t\), while not necessarily requiring smoothness over other covariates \(x\). TsBART is based on the Bayesian Additive Regression Trees (BART) model, an ensemble of regression trees. TsBART extends BART by parameterizing each tree’s terminal nodes with smooth functions of \(t\), rather than independent scalars. Like BART, tsBART captures complex nonlinear relationships and interactions among the predictors. But unlike BART, tsBART guarantees that the response surface will be smooth in the target covariate. This improves interpretability and helps regularize the estimate. After introducing and benchmarking the tsBART model, we apply it to our motivating example: pregnancy outcomes data from the National Center for Health Statistics. Our aim is to provide patient-specific estimates of stillbirth risk across gestational age (\(t\)), based on maternal and fetal risk factors (\(x\)). Obstetricians expect stillbirth risk to vary smoothly over gestational age, but not necessarily over other covariates, and tsBART has been designed precisely to reflect this structural knowledge. The results of our analysis show the clear superiority of the tsBART model for quantifying stillbirth risk, thereby providing patients and doctors with better information for managing the risk of fetal mortality. All methods described here are implemented in the R package tsbart.

**Keynote Lecture II**

**Are You Sure? The Power of Statistical Thinking**

Steve Ruberg
Analytix Thinking, LLC

The world is exploding with data, and the application of ‘analytics’ is growing at a commensurate rate. The question remains: “Are we using the right data, and are we applying smart analytics?” This talk will explore three topics that are major trends in our current scientific environment – big data, inference and validated analysis. Pros and cons of each of these will be explored, and fundamental concepts elucidated to make our statistical thinking clearer and more precise. Statisticians must grasp these trends, understand the fundamental issues underlying them and do a better job communicating how to do smart analytics. This can influence policy decisions by governments and have major societal impact, all for the betterment of mankind. This will also require statisticians to reinvent themselves to play a much broader role in the most difficult problems of our time.

**Keynote Lecture III**

**Modernizing Statistics Through Treatment Regimes: A Review**

Marie Davidian
North Carolina State University

Statistics has played and will continue to play a fundamental role in health sciences research and in particular in the design and analysis of studies of treatments for chronic diseases and disorders, and the current focus on precision medicine presents numerous opportunities for statistics. The past decade has seen an explosion of statistical methodological research for discovery of optimal treatment regimes from data; a treatment regime is a set of sequential decision rules that maps individual patient characteristics to recommended treatments from among the available, feasible options. Precision medicine seeks to make clinical decision-making evidence-based; thus, methodology for estimation of treatment regimes from data has a significant role to play in advancing this objective. This talk will review the evolution of methodology for treatment regimes and advocate for broader adoption of treatment regimes in practice.
Session 40: From bulk tissue to single cells: advances in statistical genetics and genomics

Fast and accurate alignment of single-cell RNA-seq samples using kernel density matching

Mengjie Chen, Qi Zhan and Yang Li
University of Chicago
mengjiechen@uchicago.edu

With technologies improved dramatically over recent years, single cell RNA-seq (scRNA-seq) has been transformative in studies of gene regulation, cellular differentiation, and cellular diversity. As the number of scRNA-seq datasets increases, a major challenge will be the standardization of measurements from multiple different scRNA-seq experiments enabling integrative and comparative analyses. However, scRNA-seq data can be confounded by severe batch effects and technical artifact. In addition, scRNA-seq experiments generally capture multiple cell-types with only partial overlaps across experiments making comparison and integration particularly challenging. To overcome these problems, we have developed a method, dmatch, which can both remove unwanted technical variation and assign the same cell(s) from one scRNA-seq dataset to their corresponding cell(s) in another dataset. By design, our approach can overcome compositional heterogeneity and partial overlap of cell types in scRNA-seq data. We further show that our method can align scRNA-seq data accurately across tissues biopsies.

SCOPE: A normalization and copy number estimation method for single-cell DNA sequencing

Rujin Wang, Danyu Lin and Yuchao Jiang
University of North Carolina at Chapel Hill
yuchaoj@email.unc.edu

Whole genome single-cell DNA sequencing (scDNA-seq) enables characterization of copy number profiles at the cellular level. This circumvents the averaging effects associated with bulk-tissue sequencing and has increased resolution yet decreased ambiguity in deconvolving cancer subclones and elucidating cancer evolutionary history. ScDNA-seq data is, however, sparse, noisy, and highly variable even within a homogeneous cell population, due to the biases and artifacts that are introduced during the library preparation and sequencing procedure. Here, we propose SCOPE, a normalization and copy number estimation method for scDNA-seq data. The distinguishing features of SCOPE include: (i) utilization of cell-specific Gini coefficients for quality controls and for identification of normal/diploid cells, which are further used as negative control samples in a Poisson latent factor model for normalization; (ii) modeling of GC content bias using an expectation-maximization algorithm embedded in the Poisson generalized linear model, which accounts for the different copy number states along the genome; (iii) a cross-sample iterative segmentation procedure to identify breakpoints that are shared across cells from the same genetic background. We evaluate performance of SCOPE on real scDNA-seq data sets from cancer genomic studies. Compared to existing methods, SCOPE more accurately estimates subclonal copy number aberrations and is shown to have higher correlation with array-based copy number profiles of purified bulk samples from the same patient. We further demonstrate SCOPE on three recently released data sets using the 10X Genomics single-cell CNV pipeline and show that it can reliably recover 1% of the cancer cells from a background of normal.

Benchmarking scRNA-seq clustering methods using multi-parameter ensembles of simulated data

Xianing Zheng and Jun Li
University of Michigan
junzli@med.umich.edu

In recent years, single-cell RNA sequencing (scRNA-seq) has emerged as a powerful technology for surveying cell types and state transitions in a cost-effective manner. As expected, the rapid adoption of single-cell measurement technologies has created a new pressure point around the computational analyses of these data. As of early 2019, > 350 tools have appeared to address > 30 scRNA-seq analysis tasks (e.g., normalization, clustering, and imputation). However, the community still struggles to identify the best workflow for any given task, including clustering, which is central for cell type discovery. Benchmarking studies to date have mainly relied on real datasets, tested across published methods at their default settings (e.g., Duò, A., et al. 2018). We expanded these efforts by creating an ensemble of truth-known simulated data, and testing them across many parameter combinations of existing methods. We produced scRNA-seq counts matrices for k=4 known clusters by systematically altering five parameters: cluster separation (4 levels of difficulty), cluster size (equal/unequal) and cluster distance (equal/unequal), cell library size (four levels), cell library size variability (four levels), and dropout rate (four levels), creating a full combination of 4^5 = 1,024 datasets with known statistical properties. For each dataset we evaluated the performance over 15 clustering methods and for each, adopted three gene selection methods and five k values (2,3,4,5,6), for 225 workflows, and 1024 × 225 = 230,400 runs. The performance of each run is reported by the Adjusted Rand Index and F-measure. Overall, SC3 and Seurat performed well for most of the 1024 datasets. For gene selection, highly variable and highly expressed genes gave comparable performances and are better than using highly dropout genes, which may require further in-depth tuning. Performance variation over the five parameters revealed strengths/weaknesses of individual methods/workflows. For example, RaceID2 is particularly sensitive to the true cluster distance. Seurat and SC3 are robust across varying library sizes, whereas SIMR and RaceID2 only perform well for moderate-to-high library size. The ability to benchmark algorithmic choices (e.g., selection of clustering and gene selection methods) using simulations that cover a wide parameter space has become essential in developing customized pipelines that match the inherent statistical properties in each real study.

Novel Deconvolution of Bulk Transcriptomics Data

Fei Zou1, Dong Li and Xiaojing Zheng
1University of North Carolina at Chapel Hill
feizou@email.unc.edu

Deconvolution of bulk transcriptomics data from mixed cell populations is important to identify cellular mechanism of complex diseases. In this talk, we will describe a SVA-based deconvolution method where unmeasured cell proportions will be treated as hidden variables. Top ranked surrogate variables will be selected and used to deconvolute bulk samples using cell proportions from a set of training samples or using relative cell proportions from samples with one or more missing cell types. In contrast to existing methods, the proposed approach is more flexible in adjusting effects of observed covariates, such as gender and age, and effects of hidden confounding factors whose effects, if ignored, can lead to erroneous cell proportion estimates. The advantages of the proposed method will be illustrated by both simulated and real data.
Session 41: Advances in Meta-Analysis

Meta-analysis of incidence of rare events using individual participant data
Chen Chen1, Yan Ma, Yong Ma and Qing Pan

Meta-analysis (M-A) is a quantitative approach that provides a systematic way of integrating the findings of individual research studies. Individual participant data (IPD) M-A has become an increasingly popular method in recent years. Comparing to aggregate data (AD) M-A, which only produces unadjusted estimates, IPD M-A can adjust for baseline characteristics or potential confounding variables, leading to increased power. In this study, we propose a novel method for IPD M-A of incidence rate of rare events. This exact likelihood method, built on a Poisson-Gamma hierarchical model, is considered superior to the conventional naive approaches. We also consider zero inflated models, which can be intuitively applied on rare events data. These methods are compared through a comprehensive simulation study.

A Bayesian Hierarchical Model Estimating CACE in Meta-analysis of Randomized Clinical Trials
Jincheng Zhou, James Hodges, M. Fareed Kuri and Haitao Chu

Noncompliance to assigned treatment is a common challenge in analysis and interpretation of randomized clinical trials. The complier average causal effect (CACE) approach provides a useful tool for addressing noncompliance, where CACE is defined as the average difference in potential outcomes for the response in the subpopulation of subjects who comply with their assigned treatments. In this article, we present a Bayesian hierarchical model to estimate the CACE in a meta-analysis of randomized clinical trials where compliance may be heterogeneous between studies. Between-study heterogeneity is taken into account with study-specific random effects. The results are illustrated by a re-analysis of a meta-analysis comparing the effect of epidural analgesia in labor versus no or other analgesia in labor on the outcome cesarean section, where noncompliance varied between studies. Finally, we present simulations evaluating the performance of the proposed approach and illustrate the importance of including appropriate random effects and the impact of over- and under-fitting.

On the efficiency of network meta-analysis
Lifeng Lin

Network meta-analysis (NMA) has become an increasingly used tool to compare multiple treatments simultaneously by synthesizing direct and indirect evidence in clinical research. However, the synthesized overall evidence is seldom compared with the direct evidence to validate the efficiency of an NMA. On the one hand, we propose three new measures (i.e., the effective number of studies, the effective sample size, and the effective precision) to preliminarily quantify overall evidence gained in NMAs at the pre-analysis stage. They permit evidence users to intuitively evaluate the benefit of performing NMAs, compared with pairwise meta-analyses based on only direct evidence. We use an illustrative example to demonstrate their derivations and interpretations. On the other hand, at the post-analysis stage, we use the recently proposed borrowing of strength (BoS) statistic to empirically evaluate the benefits by incorporating indirect evidence in 40 published NMAs. The BoS statistic quantifies the percentage reduction in the uncertainty of the effect estimate when adding indirect evidence to an NMA. We found that the incremental gain may reliably occur only when at least two head-to-head studies are available and treatments are well connected. Researchers should routinely report and compare the results from both network and pairwise meta-analyses.

Questionnaire on Network Meta-Analysis to Assess Its Relevance and Credibility
Joseph Cappelleri

A task force from the International Society for Pharmacoeconomics and Outcomes Research developed a consensus-based 26-item questionnaire to help decision makers assess the relevance and credibility of network meta-analysis to help inform health-care decision making. The relevance domain of the questionnaire (4 questions) calls for assessments about the applicability of network meta-analysis results to the setting of interest to the decision maker. The remaining 22 questions belong to an overall credibility domain and pertain to assessments about whether the network meta-analysis results provide a valid answer to the question they are designed to answer by examining 1) the evidence base used, 2) analysis methods, 3) reporting quality and transparency, 4) interpretation of findings, and 5) conflicts of interest. The questionnaire aims to help readers of network meta-analysis opine about their confidence in the credibility and applicability of the results of a network meta-analysis, and help make decision makers aware of the subtleties involved in the analysis of networks of randomized trial evidence. In this presentation, these recommended principles on relevance and credibility of network meta-analysis will be explained.

Session 42: Advance in Statistical Methods for Large and Complex Data

Comparing and weighting imperfect models using D-probabilities
Meng Li1 and David Dunson2

The Bayesian paradigm provides a natural way to deal with uncertainty in model selection by assigning each model in a list of models under consideration a posterior probability. Unfortunately, this framework relies on the assumption that one of the models in the list is the true model. When this assumption is violated and all the models are imperfect, interpretation of posterior model probabilities is unclear. We propose a new concept of absolute model probabilities, which measure the quality of imperfect models. This concept leads to divergence-based estimates D-probabilities relying on evaluating parametric models relative to a nonparametric Bayesian reference using Kullback-Leibler divergence. While providing goodness-of-fit assessment, D-probabilities avoid some of the pitfalls of usual posterior model probabilities including large sensitivity to prior choice. In an application to linear model selection against a Gaussian process reference, we provide simple analytic forms for routine implementation and show that D-probabilities automatically penalize model complexity. Some asymptotic properties of this framework are described. Absolute model probabilities have several interesting probabilistic interpretations, and can potentially be applied in broad problems.
A Sparse Random Projection-based Test for Overall Qualitative Treatment Effects

Chengchun Shi, Wenbin Lu and Rai Song
North Carolina State University

cshi4@ncsu.edu

In contrast to the classical "one-size-fits-all" approach, precision medicine proposes the customization of individualized treatment regimes to account for patients' heterogeneity in response to treatments. Most of existing works in the literature focused on estimating optimal individualized treatment regimes. However, there has been less attention devoted to hypothesis testing regarding the existence of overall qualitative treatment effects, especially when there is a large number of prognostic covariates. When covariates don’t have qualitative treatment effects, the optimal treatment regime will assign the same treatment to all patients regardless of their covariate values. In this paper, we consider testing the overall qualitative treatment effects of patients’ prognostic covariates in a high dimensional setting. We propose a sample splitting method to construct the test statistic, based on a nonparametric estimator of the contrast function. When the dimension of covariates is large, we construct the test based on sparse random projections of covariates into a low-dimensional space. We prove the consistency of our test statistic. In the regular cases, we show the asymptotic power function of our test statistic is asymptotically the same as the "oracle" test statistic which is constructed based on the "optimal" projection matrix. Simulation studies and real data applications validate our theoretical findings.

Estimating the causal effect of treatment regimes for organ transplantation

David Vock and Jeffrey Boatman
University of Minnesota

tyock@umn.edu

Patients awaiting cadaveric organ transplantation face a difficult decision if offered a low-quality organ: accept the organ or remain on the waiting list and hope a better organ is offered in the future. A dynamic treatment regime (DTR) for transplantation is a rule that determines whether a patient should decline an offered organ. Existing methods can estimate the effect of DTRs on survival outcomes, but these were developed for applications where treatment is abundantly available. For transplantation, organ availability is limited, and existing methods can only estimate the effect of a DTR assuming a single patient follows the DTR. We show for transplantation that the effect of a DTR depends on whether other patients follow the DTR. To estimate the anticipated survival if the entire population awaiting transplantation were to adopt a DTR, we develop a novel inverse probability weighted estimator (IPCW) which re-weights patients based on the probability of following their transplant history in the counterfactual world in which all patients follow the DTR of interest. We estimate this counterfactual probability using hot deck imputation to fill in data that is not observed for patients who are artificially censored by IPCW once they no longer follow the DTR of interest. We show via simulation that our proposed method has good finite-sample properties, and we apply our method to a lung transplantation observational registry.

A general Classification framework to estimate the optimal dynamic treatment regime

Baqun Zhang and Min Zhang

1 Shanghai University of Finance and Economics
2 University of Michigan

A dynamic treatment regime is a sequence of decision rules, each corresponding to a decision point, that determine that next treatment based on each individual’s own available characteristics and treatment history up to that point. We show that identifying the optimal dynamic treatment regime can be recast as a sequential optimization problem and propose a direct sequential optimization method to estimate the optimal treatment regimes. In particular, at each decision point, the optimization is equivalent to sequentially minimizing a weighted expected misclassification error. Based on this classification perspective, we propose a powerful and flexible C-learning algorithm to learn the optimal dynamic treatment regimes backward sequentially from the last stage until the first stage. C-learning is a direct optimization method that directly targets optimizing decision rules by exploiting powerful optimization/classification techniques and it allows incorporation of patient’s characteristics and treatment history to improve performance, hence enjoying advantages of both the traditional outcome regression based methods (Q- and A-learning) and the more recent direct optimization methods. The superior performance and flexibility of the proposed methods are illustrated through extensive simulation studies.

Session 43: New Methods for Survival Analysis and Network Analysis with Application to Biomarker Studies

Learning causal networks via additive faithfulness

Kuang-Yao Lee, Tianqi Liu, Bing Li and Hongyu Zhao

1 Temple University
2 Google, LLC
3 Pennsylvania State University
4 Yale University

kuangyao.l@gmail.com

In this paper 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 multi-dimensional kernels. This distinct feature strikes a balance between a parametric model and a fully nonparametric model, which makes the proposed model attractive for handling large networks. We develop an estimator for AFDAG based on a linear operator that characterizes ACI, and establish the consistency and convergence rates of this estimator, as well as the uniform consistency of the estimated DAG. Moreover, we introduce a modified PC-algorithm to implement the estimating procedure efficiently, so that its complexity is determined by the level of sparseness 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.

Identifying Disease-Associated Biomarker Network Features Through Conditional Graphical Model

Shanghong Xie, Donglin Zeng and Yuanjia Wang

1 Columbia University
2 University of North Carolina at Chapel Hill

sx2156@columbia.edu

Biomarkers are often organized into networks, in which the strengths of network connections vary across subjects depending on...
subject-specific covariates (e.g., genetic variants). Variation of network connections, as subject-specific feature variables, has been found to predict disease clinical outcome. In this work, we develop a two-stage method to estimate biomarker networks that account for heterogeneity among subjects and evaluate networks association with disease clinical outcome. In the first stage, we propose a conditional Gaussian graphical model with mean and precision matrix depending on covariates to obtain subject-specific networks. In the second stage, we evaluate clinical utility of network measures (connection strengths) estimated from the first stage. The second stage analysis provides the relative predictive power of between-region network measures on clinical impairment in the context of within-region biomarkers and existing disease risk factors. We assess the performance of proposed method by extensive simulation studies and application to a Huntingtons disease (HD) study to investigate the effect of HD causal gene on the rate of change in motor symptom through affecting brain subcortical and cortical grey matter atrophy connections. We show that cortical network connections and subcortical volumes, not but subcortical connections are identified to be predictive of clinical motor function deterioration. We validate these findings in an independent HD study. Lastly, highly similar patterns seen in the grey matter connections and a previous white matter connectivity study suggest a shared biological mechanism for HD and support the hypothesis that white matter loss is a direct result of neuronal loss as opposed to the loss of myelin or dysmyelination.

**Generalized Mean Residual Life Models for Case-Cohort and Nested Case-Control Studies**

Peng Jin, Anne Zeleniuch-Jacquotte and Mengling Liu
New York University

Mean residual life (MRL) is defined as the remaining life expectancy of a subject who has survived to a certain time point, and is an alternative way to hazard function for characterizing time-to-event data. Inference and application of MRL models have primarily focused on cohort studies. In practice, case-cohort and nested case-control designs have been commonly used within large cohort studies, particularly when studying costly molecular biomarkers. They enable prospective inference as the full-cohort design with significant cost-saving benefit. In this paper, we study the modeling and inference of a family of generalized MRL models for studies under case-cohort and nested case-control designs. Built upon the idea of inverse selection probability, the weighted estimating equations are constructed to estimate the regression parameters and baseline mean residual life function. Asymptotic properties of the proposed estimators are established and finite-sample properties are evaluated by extensive numerical simulations under various settings. The proposed methodology is demonstrated on the NYU Women’s Health Study.

**Adjusting for Handling Effects in Microarray Data for Survival Risk Prediction**

Andy Ni1, Mengling Liu2 and Li-Xuan Qin3
1Ohio State University
2New York University
3Memorial Sloan Kettering Cancer Center

Survival risk prediction is an important task in clinical cancer research. By its virtue of simultaneously measuring the expression of thousands of genes, microarrays have been a highly useful tool for the development of survival risk models based on patients’ gene expression profiles. A well-known caveat of microarray measurements, however, is that they are often contaminated with handling effects arising from non-uniform experimental handling due to, for example, batch effects and technician effects. It has been shown in the context of group comparison and classification that the negative impact of handling effects can be prevented effectively by careful study design and moderately by post-hoc data normalization. Research is still lacking on how handling effects impact survival risk prediction. In this study, we conducted extensive simulations based on a unique pair of microRNA datasets on ovarian cancer from Memorial Sloan Kettering Cancer Center to elucidate this impact under both univariate analysis and multivariate regularized Cox regression. To preserve the correlation structure of the expression profile and the distribution of the progression-free survival (PFS) in the real data, we developed a permutation-based technique to simulate data with desired association between microRNA expression and PFS. We proposed a stratification strategy in survival modeling to reduce the impact of handling effects and compared its prediction accuracy with that of data normalization. Our results suggest that stratification consistently outperforms normalization over variation scenarios of training-data handling effects and test-data handling effects and choice of survival models.

**Session 44: Statistical Challenges at the Intersection of Prediction and Causality**

**Causal inference in algorithmic fairness**

Joshua Loftus1, Matt Kasner2, Ricardo Silva3 and Chris Russell1
1New York University
2Cambridge University
3University College London

In this talk I will survey some recent literature on algorithmic fairness with a focus on methods based on causal inference. One such approach, counterfactual fairness, requires that predictions or decisions should be the same both in the actual world and in a counterfactual world where an individual had a different value of a sensitive attribute, such as race or gender. This approach defines fairness in the context of a causal model for the data which is also useful for thinking about the implicit assumptions or consequences of other definitions, or designing interventions.

**The Certification Framework**

Jordan Rodu1 and Michael Baiocchi2
1University of Virginia
2Stanford University

The use of predictive algorithms to generate actionable insights has reshaped the landscape of business, communication, government, and science. We are interested in the question of how a data analyst thinking about the implicit assumptions or consequences of other definitions, or designing interventions.

**Targeted Smooth Bayesian Causal Forests for Heterogeneous Time-varying Treatment Effects**

Jennifer Starling, Jared Murray, Carlos Carvalho, Radek Bukowski and James Scott
University of Texas at Austin
Bayesian Additive Regression Trees (BART) has been shown to be an effective framework for modeling nonlinear regression functions, with strong predictive performance in a variety of contexts. The BART prior over a regression function is defined by independent prior distributions on tree structure and leaf or end-node parameters. We consider two flavors of BART. First, BART with Targeted Smoothing induces smoothness over a single covariate by replacing independent Gaussian leaf priors with smooth functions. Second, Bayesian Causal Forests (BCF) has successfully adapted BART for estimating heterogeneous treatment effects from observational data, particularly in cases where standard methods yield biased estimates due to strong confounding. BCF parameterizes BART models to allow for separate regularization of the treatment and prognostic effects, making it possible to shrink towards homogeneity. We introduce a new version of the BCF prior which incorporates targeted smoothing for modeling heterogeneous treatment effects which vary smoothly over a target covariate. We demonstrate the utility of this approach by applying our model to a timely women’s health problem: comparing two dosing regimens for an early medical abortion protocol, where the outcome of interest is the probability of a successful early medical abortion procedure at varying gestational ages, conditional on patient covariates. We discuss the benefits of this approach in a variety of other women’s health and obstetrics modeling problems where gestational age is a typical covariate, as well as potential applications in a range of other disciplines.

Estimating heterogeneous treatment effects in the presence of unobserved confounding: hybridizing an instrumental variable study design and with a machine learning approach

Michael Baiocchi
Stanford University
mike.baiocchi@gmail.com

If we want to identify subgroup heterogeneity in response to interventions then we will need to use large, observational databases because it is often too expensive to deploy high-quality, prospectively design data from randomized controlled trials. Leveraging prior work on nonparametric estimation of responses in the presence of unobserved confounding, we propose a novel hybridized approach of instrumental variable study design with machine learning based detection step. This talk is designed to be accessible to non-experts, building up motivation from an applied health care example.

Session 45: RWE in Drug Development and Observational Studies: Examples and Discussion

Outcome weighted learning for imbalanced high dimensional precision medicine problems

Hui Sun1, Bruce Craig and Lingsong Zhang
Purdue University
sunn400@purdue.edu

Classification has broad application in precision medicine. Single-stage precision medicine problems can be reformulated as weighted classification problems. The subtle differences between classification methods may lead to different application performances under high dimension, low sample size (HDLSS) setting. Among the margin-based classification methods, we propose the distance weighted discrimination outcome weighted learning method (DWD-OWL) as an alternative to outcome weighted learning with hinge loss. We show through simulation that the DWD-OWL will outperform SVM-OWL. The proofs of Fisher consistency for DWD-OWL in both the binary and multicategory cases are provided. Under mild conditions, the insensitivity of DWD-OWL for imbalanced data is also demonstrated.

Set-valued Classification and Its Applications to Precision Medicine

Xingye Qiao
Binghamton University
qiao@math.binghamton.edu

In this talk, I will discuss the practical usefulness for set-valued classification, and introduce two frameworks to achieve it, namely, classification with reject and refine options, and confidence set learning. Then I will discuss an application of the the set-valued classification to the precision medicine area, specifically, to individualized treatment recommendations.

Deep Learning in Medical Image Recognition: Diabetic Retinopathy (DR) Diagnosis

Xuanyao He1 and Faustine Li
1Eli Lilly and Company
he_xuanyao@lilly.com

Diabetic Retinopathy (DR) is a common disease of the retina among patients with diabetes mellitus and can be 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 manual screenings. Many areas and countries are suffering from lack of professionals to diagnose DR for the big population. Automatic screening methods of DR using images with high accuracy have the potential to assist physicians in evaluating patients earlier, thereby potentially enabling them to seek timely preventions and treatments. These methods emphasize on DR diagnosis using appropriate image processing and data mining techniques. In particular, in our 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 learning model, fine tuning and neural networks. Examples will be provided to demonstrate the ability of machine learning techniques to help solve this kind of important medical problem.

Target population statistical inference with data integration across multiple sources

Yang Song1 and Xihao Li2
1Vertex Pharmaceuticals, Inc.
2Harvard University
yang_song@vrtx.com

For clinical trials in rare disease areas, a major challenge is the limited amount of available information for making robust statistical inference due to low disease prevalence. External data sources present data integration opportunities to enhance statistical inference. We propose an intuitive integrated inference framework to integrate information from all relevant data sources and make inference on the treatment effect over a specific target population. The method is easily implemented and extended with modern machine learning tools. It is complemented by a variance estimation procedure to facilitate statistical inference. The proposed method is shown to have good statistical properties with both theoretical development and simulation studies. We argue that the integrated inference framework not only provides an intuitive and coherent perspective for a wide range of clinical trial inference problems but also has broad application areas in clinical trial settings and beyond, as a quantitative data integration tool for making robust inference in a target population precise manner for policy and decision makers.

Deep Learning in Medical Image Recognition: Diabetic Retinopathy (DR) Diagnosis

Xuanyao He1 and Faustine Li
1Eli Lilly and Company
he_xuanyao@lilly.com

Diabetic Retinopathy (DR) is a common disease of the retina among patients with diabetes mellitus and can be 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 manual screenings. Many areas and countries are suffering from lack of professionals to diagnose DR for the big population. Automatic screening methods of DR using images with high accuracy have the potential to assist physicians in evaluating patients earlier, thereby potentially enabling them to seek timely preventions and treatments. These methods emphasize on DR diagnosis using appropriate image processing and data mining techniques. In particular, in our 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 learning model, fine tuning and neural networks. Examples will be provided to demonstrate the ability of machine learning techniques to help solve this kind of important medical problem.
Estimation of a functional of high-dimensional mean and covariance matrix

Haolei Weng\textsuperscript{1}, Yifeng Zhou\textsuperscript{2} and Jianqing Fan\textsuperscript{2}
\textsuperscript{1}Michigan State University
\textsuperscript{2}Princeton University

wenghaol@msu.edu

We consider estimating a functional involving both the mean vector and covariance matrix. The functional carries important information across a sequel of multivariate statistics problems. We study the minimax estimation of the functional in the high-dimensional setting. Akin to past works on functional estimation, we reveal that the minimax estimation rate of the functional undergoes a phase transition between regular parametric rate and certain form of high-dimensional rate. In addition, we show how the optimal rate is attained by a carefully designed plug-in estimator based on de-biasing. On the contrary, a family of naive plug-in estimators are proved to fall short.

Structured Latent Factor Analysis for Large-scale Data

Xiaou Li\textsuperscript{1}, Yunxia Chen\textsuperscript{2} and Siliang Zhang\textsuperscript{3}
\textsuperscript{1}University of Minnesota
\textsuperscript{2}London School of Economics and Politics
\textsuperscript{3}Pudan University

lixx1766@umn.edu

Latent factor models are widely used to measure unobserved latent traits in social and behavioral sciences, including psychology, education, and marketing. When used in a confirmatory manner, design information is incorporated, yielding structured latent factor models. Motivated by the applications of latent factor models to large-scale measurements which consist of many manifest variables and a large sample size, we study the properties of structured latent factor models under an asymptotic setting where both the number of manifest variables and the sample size grow to infinity. We establish necessary and sufficient conditions on the measurement design that ensure the structural identifiability under a general family of structured latent factor models. In addition, we propose an estimator that can consistently recover the latent factors under mild conditions.

Latent factor models are widely used to measure unobserved latent traits in social and behavioral sciences, including psychology, education, and marketing. When used in a confirmatory manner, design information is incorporated, yielding structured latent factor models. Motivated by the applications of latent factor models to large-scale measurements which consist of many manifest variables and a large sample size, we study the properties of structured latent factor models under an asymptotic setting where both the number of manifest variables and the sample size grow to infinity. We establish necessary and sufficient conditions on the measurement design that ensure the structural identifiability under a general family of structured latent factor models. In addition, we propose an estimator that can consistently recover the latent factors under mild conditions.

Bridging the gap between noisy healthcare data and knowledge: causality and portability

Xu Shi\textsuperscript{1}, Xiaou Li\textsuperscript{2} and Tianxi Cai\textsuperscript{2}
\textsuperscript{1}Harvard University
\textsuperscript{2}University of Minnesota

shixu0830@gmail.com

Routinely collected healthcare data present numerous opportunities for biomedical research but also come with unique challenges. For example, critical issues such as data quality, unmeasured and mismeasured confounding, high-dimensional covariates, and patient privacy concerns naturally arise. In this talk, I present tailored causal inference methods and automated data quality control pipeline that aim to overcome these challenges and make the transition from data to knowledge. I detail the challenge of inconsistent “languages” used by different healthcare systems and coding systems. In particular, different healthcare providers may use alternative medical codes to record the same diagnosis or procedure, limiting the transportability of phenotyping algorithms and statistical models across healthcare systems. I formulate the idea of medical code translation into a statistical problem of inferring a mapping between two sets of multivariate, unit-length vectors learned from two healthcare systems, respectively. The statistical problem is particularly interesting because the training data is corrupted by a fraction of mismatch in the response-predictor pairs, whereas classical regression analysis tacitly assumes that the response and predictor are correctly linked. I propose a novel method for mapping recovery and establish theoretical guarantees for estimation and model selection consistency.

Structured Latent Factor Analysis for Large-scale Data

Xiaou Li\textsuperscript{1}, Yunxia Chen\textsuperscript{2} and Siliang Zhang\textsuperscript{3}
\textsuperscript{1}University of Minnesota
\textsuperscript{2}London School of Economics and Politics
\textsuperscript{3}Pudan University

lixx1766@umn.edu

Latent factor models are widely used to measure unobserved latent traits in social and behavioral sciences, including psychology, education, and marketing. When used in a confirmatory manner, design information is incorporated, yielding structured latent factor models. Motivated by the applications of latent factor models to large-scale measurements which consist of many manifest variables and a large sample size, we study the properties of structured latent factor models under an asymptotic setting where both the number of manifest variables and the sample size grow to infinity. We establish necessary and sufficient conditions on the measurement design that ensure the structural identifiability under a general family of structured latent factor models. In addition, we propose an estimator that can consistently recover the latent factors under mild conditions.

Characterization of heatwaves is becoming increasingly important in environmental research as they pose a significant threat to many living organisms around the world. Though several quantification of the extremities of ambient temperature and relative humidity have been proposed to characterize heatwaves, they are mostly improved and there does not exist a broadly accepted definition of heatwave. The proposed work is based on a general probabilistic approach based on determining the probability distributions of the duration of chosen meteorological variables (e.g., ambient temperature etc.) which can capture the essence of all existing ad-hoc existing definitions of heatwaves. An exact distribution theory is obtained for frequency distribution of the duration for a stationary Markov process, and also an approximate distribution of duration is obtained under more general dependent sequence of meteorological variables. For a given site, using a daily time series (of ambient temperature or heat-index) we define a heatwave as the number of sustained days above a given threshold using the quantiles of probability distribution of the duration. We illustrate the proposed methodology using daily time series of ambient temperature for a fixed site (e.g., Atlanta). Further, the methodology is extend to derive spatially smoothed version of the probability distributions and is illustrated using the USCRN consisting of 126 sites across the U.S.
Argo floats measure seawater temperature and salinity in the upper 2000 meters of the ocean. These floats are uniquely capable of measuring the heat content of the global ocean, a quantity that is of central importance for understanding changes in the Earth’s climate system. But providing detailed spatio-temporal estimates of the heat content is statistically challenging due to the complex structure and large size of the Argo dataset. We have previously demonstrated (Kuusela and Stein, 2018) that locally stationary Gaussian process regression leads to improved and computationally efficient interpolation of Argo data. Here we build upon those findings to produce improved Argo-based global ocean heat content estimates. We study the sensitivity of these estimates to the underlying statistical assumptions and present results indicating that the magnitude of the overall warming trend may depend on the modeling of the climatological time trend in the mean field estimate. We also investigate the benefits of including time in the interpolation and propose a method for uncertainty quantification that yields appropriate spatial correlations without the need for a global covariance model.

**Session 48: Real world evidence and big data**

**Globally scalable quantile regression with SGD for massive real-world data**

Yixin Fang1, Jinfeng Xu2 and Na Zhao2

1 Abbvie, Inc.
2 University of Hong Kong

We consider quantile regression for large-scale or streaming data where it is numerically challenging or sometimes infeasible to store the entire dataset in memory. In such situations, stochastic gradient descent methods that process one data point at a time and recursively estimate the parameters are desirable, exhibiting both numerical convenience and memory efficiency. The development of stochastic gradient descent methodology for quantile regression is most relevant when the covariate-response association is examined globally over a continuum set of quantile levels rather than locally at a single level. For globally concerned quantile regression, the estimand is a functional instead of a finite-dimensional vector, which motivates us to propose a novel approach for recursive estimation via stochastic gradient descent. To quantify the uncertainty associated with the estimated regression quantiles and facilitate statistical inferences, we further develop a random perturbation-based resampling strategy that specifically caters to the large-scale or online setting. We also establish some theoretical results such as uniform consistency and weak convergence.

**Analytic Strategies for Economic Endpoints in Real-World Studies**

Hongwei Wang, Yixin Fang and Weili He

Abbvie, Inc.

Real-world studies play an essential role in characterizing an intervention’s effectiveness and safety in routine clinical practice in terms of clinical, economic and humanistic outcomes. The economic endpoints usually take the form of direct cost, such as healthcare resource utilization (HCRU) and associated expenditures, and indirect cost, such as work productivity and burden to caregiver. For group comparison, in addition to the confounding and bias that are inherent in real-world studies, the economic endpoints exhibit several properties that need to be further carefully considered at the analytic stage: 1) not every patient incurs HCRU, e.g., hospitalization,
which resulting in substantial proportion of zeros; 2) the distribution of cost data is highly skewed and existence of extreme values may arise; 3) it is a function of study time period which generally varies from one patient to another. Built upon the rich literatures, we will review several analytic strategies, including ordinary linear regression, lognormal regression, generalized linear regression assuming various distribution, two-stage Tobit model, regression tree and Cox proportional hazard model, and offer practical consideration for their applications.

Survival Extrapolation in Cost-effectiveness Analysis: Combining Clinical Trial and Real-World Data
Bining Yu
AstraZeneca

Most cost-effectiveness analyses (CEA) are based on evidence from randomized clinical trials which usually do not follow all patients until death. Extrapolation of survival beyond the follow-up period is critical for reliable estimates of life expectancy, and incremental cost-effectiveness ratio (ICER). Evidence synthesis is a valuable approach that combines clinical trial data, expert opinions, and real-world data from population-based databases and electronic health records. We describe the implementation of mixture cure model and Bayesian method in the survival extrapolation. We discuss the advantage and possible pitfalls of survival extrapolations and give an example of cost-effectiveness analysis of an immuno-oncology therapy.

Session 49: Statistical Learning Advancement for Inference in Big Data Age

Statistical Inference on Partially Linear Panel Model under Unobserved Linearity
† Ruiqi Liu, Ben Boukai and Zuofeng Shang
Indiana University-Purdue University Indianapolis
rliu4@binghamton.edu

A new statistical procedure, based on a modified spline basis, is proposed to identify the linear components in the panel data model with fixed effects. Under some mild assumptions, the proposed procedure is shown to consistently estimate the underlying regression function, correctly select the linear components, and effectively conduct the statistical inference. When compared to existing methods for detection of linearity in panel model, our approach is demonstrated to be theoretically justified as well as practically convenient. We provide a computational algorithm which implements the proposed procedure along with a path-based solution method for linearity detection, which avoids the burden of selecting the tuning parameter for the penalty term. Monte Carlo simulations are conducted to examine the finite sample performance of our proposed procedure with detailed findings that confirm our theoretical results in the paper. Application to Aggregate Production Data also illustrates the necessity for detecting linearity in the partially linear panel model.

Asymptotic Properties of Neural Network Sieve Estimators
× Xiaoxi Shen, Chang Jiang, Lyudmila Sakhanenko and Qing Lu
Michigan State University
xiaoximichael@gmail.com

Neural networks have been widely used in machine learning and artificial intelligence nowadays. Due to the universal approximation theorem (Hornik, Stinchcombe and White, 1989), a neural network with one hidden layer can approximate any continuous functions on compact supports. Statistically, a neural network is a nonlinear regression problem. However, if we considered it parametrically, due to the unidentifiability of the parameters, it is difficult to derive its asymptotic properties. Instead, we considered the estimation problem as a nonparametric regression problem and used the results from sieve estimation to establish the consistency, rates of convergence and asymptotic normality of the neural network estimators. We also investigate the validity of the theories based on some simulation results.

Auto-encoding graph-valued data with applications to brain connectomes
× Meimei Liu1, Zhengwu Zhang2 and David Dunson1
1Duke University
2University of Rochester
liu1197@purdue.edu

Our interest focuses on developing statistical methods for analysis of brain connectome graphs. Nodes in the brain graph correspond to different regions of interest (ROIs) while edges correspond to structural connections between these ROIs. Due to the high-dimensional replicated graph structure of the data, it becomes challenging to conduct analyses relating brain connectomes to other factors. Current approaches focus on summarizing the graph using either pre-specified topological features or tensor principal components analysis. In this work, we instead develop a nonlinear latent factor analysis approach for summarizing the brain graph in both unsupervised and supervised settings. The proposed approach builds on methods for hierarchical modeling of replicated graph data, as well as variational auto-encoders that use neural networks for dimensionality reduction. Hence, we refer to the method as Graph AuTo-Encoding (GATE). An efficient computational algorithm is developed, and GATE is compared with tensor PCA and other competitors through simulations and applications to data from the Human Connectome Project (HCP).

An Fast Algorithm For The A-optimal Sampling Distributions In A Big Data Linear Regression
× Hansxiang Peng, Fei Tan, Sheng Zhang and Haixia Li
Math Sciences, IUPUI
hanxpeng@iupui.edu

While classic methods compute the least squares estimate in O(np) time for sample size n and parameter dimension p, randomized subsampling methods usually take o(np^2) time. Typically, the computational bottleneck is computing the appropriate non-uniform importance sampling distribution. In this talk, we construct a randomized algorithm to compute the A-optimal sampling distribution given Peng and Tan (2018). The algorithm exploits Hadamard matrices which encode the discrete Fourier transformation over the additive group. It has the running time o(np^2). Extensive simulations and a real data application in big sample sizes will be reported about the running time and efficiency of the algorithm, with comparison to the uniform sampling.

Session 50: Moving towards Multimodal and Multi-layer imaging analysis

Joint and Individual Non-Gaussian Component Analysis for Data Integration
× Benjamin Risk1 and Irina Gaynanova2
1Emory University
2Texas A&M University
benjamin.risk@emory.edu
As advances in technology allow the acquisition of complementary information, it is increasingly common for scientific studies to collect multiple data sets. Large-scale neuroimaging studies often include multiple modalities (e.g., functional MRI, diffusion MRI, and/or structural MRI) and/or imaging data. With the aim to understand the relationships between data sets, classical approaches to data integration utilize transformations that maximize covariance or correlation. In neuroimaging, a popular approach uses principal component analysis for dimension reduction prior to feature extraction via a joint independent component analysis (ICA). We introduce Joint and Individual Non-Gaussian component analysis (JIN) for data integration, where dimension reduction and feature extraction are achieved simultaneously. We focus on information shared in subject score subspaces estimated using the Jarque-Bera statistic. We apply our method to data from the Human Connectome Project.

**Methods for Analyzing Multi-modal Imaging Data**

Kristin Linn¹, Alessandra Valcarcel², Simon Vandekar³, Theodore Satterthwaite¹ and Russell Shinohara¹
¹University of Pennsylvania
²Vanderbilt University

Local cortical coupling is a subject-specific measure of the spatially varying relationship between cortical thickness and sulcal depth. Although it is a promising first step towards understanding local covariance patterns between two image-derived measurements, a more general coupling framework that can accommodate multiple volumetric imaging modalities is warranted. We first introduce Inter-Modal Coupling (IMC), an analogue of local coupling in volumetric space that can be used to produce subject-level, spatially varying feature maps derived from two volumetric imaging modalities. We then leverage IMC to address partial volume effects when studying localized relationships between gray matter density and cerebral blood flow (CBF) among participants in the Philadelphia Neurodevelopmental Cohort. We also define population-level analogues of IMC parameters and estimate them using generalized estimating equations.

**Characterizing the Longitudinal Behavior of Multiple Sclerosis Lesions on Structural Magnetic Resonan**

Elizabeth Sweeney¹, Russell Shinohara², Blake Dewey³, John Muschelli⁴, Daniel Reich⁴, Ciprian Crainiceanu³ and Ani Eloyan¹
¹Weill School of Medicine, Cornell University
²University of Pennsylvania
³Johns Hopkins University
⁴National Institute of Neurological Disorder and Stroke

Brain structural magnetic resonance imaging (sMRI) is a tool that uses a magnetic field to produce detailed images of the brain. Patients with multiple sclerosis (MS) have lesions in their brains which are visible on sMRI. MS lesions formation is a complex process involving inflammation, tissue damage, and tissue repair - all of which are visible on sMRI and potentially modifiable by pharmacological therapy. We introduce a PCA regression modeling framework for relating voxel-level (or three-dimensional pixel-level), longitudinal, multi-sequence sMRI intensities within MS lesions to clinical information and therapeutic interventions. To do so, we first characterize the post-lesion incidence repair process on longitudinal, multi-sequence sMRI as voxel-level intensity profiles. We then perform PCA on the intensity profiles to develop a voxel-level biomarker for identifying slow and persistent, long-term intensity change within lesion voxels. We then relate the biomarker to the clinical information in a mixed model framework. Treatment with disease-modifying therapies and steroids are both associated with return of a voxel to an intensity value closer to that of normal-appearing tissue. This modeling frameworks show promise for understanding the mechanisms of tissue damage in MS and for evaluating the impact of treatments for the disease in clinical trials.

**On testing for spatial correspondence between maps of human brain structure and function**

Simon Vandekar

A critical issue in many neuroimaging studies is the comparison between brain maps. Nonetheless, it remains unclear how one should test hypotheses focused on the overlap or spatial correspondence between two or more brain maps. This “correspondence problem” affects, for example, the interpretation of comparisons between task-based patterns of functional activation, resting-state networks or modules, and neuroanatomical landmarks. To date, this problem has been addressed with remarkable variability in terms of methodological approaches and statistical rigor. In this paper, we address the correspondence problem using a spatial permutation framework to generate null models of overlap by applying random rotations to spherical representations of the cortical surface, an approach for which we also provide a theoretical statistical foundation. We use this method to derive clusters of cognitive functions that are correlated in terms of their functional neuroanatomical substrates. In addition, using publicly available data, we formally demonstrate the correspondence between maps of task-based functional activity, resting-state fMRI networks and gyral-based anatomical landmarks. We provide open-access code to implement the methods presented for two commonly-used tools for surface based cortical analysis (https://www.github.com/spin-test). This spatial permutation approach constitutes a useful advance over widely-used methods for the comparison of cortical maps, thereby opening new possibilities for the integration of diverse neuroimaging data.

**Session 51: Recent Development of Artificial Intelligence in Healthcare**

**Bayesian scalar-on-image neural network with application to skin cancer images**

Jian Kang

University of Michigan

We develop a new nonlinear scalar-on-image regression model with scalar as an outcome variable and images as predictors. We are interested in identifying important image features (pixels/voxels) that are highly predictive of the outcome variable. We construct a novel one-hidden-layer Bayesian neural network with spatially-varying coefficients (SVC). We assign soft-thresholded Gaussian process coefficients (SVC). We assign soft-thresholded Gaussian process structures, and achieving a high prediction accuracy, compared with other state-of-the-art deep learning methods. Our BNN-STGP enjoys simple network structure and has a better model interpretation.
We illustrate our methods via extensive simulation studies and an analysis of melanoma cancer images.

AI/ML for Pharma R&D: Analytical Challenges and Opportunities
Ray Liu
Takeda Pharmaceuticals
Ray.Liu@Takeda.com
Pharma R&D can use a boost to improve efficiency. With the success in other industries, AI/ML is seen as a game changer to drug development. But unlike in consumer spaces where data are abundant and cheap to generate, pharma R&D has its own constrains that could prevent the full potentials of AI/ML to be fulfilled. In the presentation we will explain the challenges faced for AI/ML implementation in drug development, and discuss right opportunities to AI/ML to shine.

Statistical Evaluation of Medical Devices that Incorporate Artificial Intelligence
Yaii Xu, Xiaobo Xiong and Changhong Song
Center for Devices and Radiological Health, FDA
Yaii.Xu@FDA.HHS.GOV
Artificial intelligence (AI) and machine learning (ML) algorithms have been incorporated in more and more applications of medical devices. Yet these devices nonetheless provide new challenges to statisticians. Recently, AI/ML related devices have been approved/cleared in multiple areas such as radiology imaging, ophthalmology, etc. Different guidance documents have also been published for FDA’s recommendations in regulating AI/ML-related medical devices in various areas. This presentation will review examples of recently approved AI/ML-based medical products and statistical methods that support the validation of medical devices for in-vitro diagnostics.

Responder Analysis by Deep Learning with Application to a Phase III Clinical Trial
Wodan Ling1, Qi Tang2 and Youran Qi3
1Columbia University
2Sanofi
3University of Wisconsin-Madison
wdling666@gmail.com
Responder analysis is routinely conducted in clinical trials to examine the treatment effect of an investigational product or medical practice. Its response is the probability that a patient is a responder, i.e. the chance that one has a response exceeding a predefined threshold. This analysis provides clinically meaningful results and may complement findings on the original continuous scale of the endpoint. To estimate the probabilities, the continuous response variable is often discretized into a categorical variable. Generalized linear models are then used to model the relationship between covariates and the discretized response. However, it may suffer from the violation of the linearity assumption. To overcome such a limitation, we propose a deep responder analysis method, which takes a deep learning approach without making a linearity assumption and variates and the discretized response. However, it may suffer from the violation of the linearity assumption. To overcome such a limitation, we propose a deep responder analysis method, which takes a deep learning approach without making a linearity assumption and offers automatic feature representation embedded. However, discretization of a continuous response variable was adopted in this method, which may lead to loss of power. Thus, in an attempt to improve the method, a second method is proposed, which avoids discretization while still enabling inference on probabilities of being a responder. The second method may be superior to the first one if there is no measurement error in the response variable and an appropriate loss function is chosen. We illustrate the two methods using a clinical trial example in the area of asthma.

Session 52: Contributed Session 1
Checking Adequacy of Variance Function with Unknown Mean Function
Jia Liang and Weixing Song
Kansas State University
jliang2019@ksu.edu
In this talk, we propose a class of test procedures to check the adequacy of parametric forms for the variance function in regression models when the mean regression function is unknown. After replacing the unknown regression function with the Nadaraya-Watson estimator, a class of test statistics are constructed based on a minimized distance between a non-parametric estimate and a parametric estimate under the null hypothesis for the variance function. The large sample properties of the minimum distance estimate for the parameters presented in the variance function are discussed, and the asymptotic distribution of the test statistics under the null hypothesis is established. Extensive simulation studies indicate that the finite sample performance of the proposed test procedure is satisfactory.

A new gene-based association test
Zhongxue Chen1 and Kai Wang2
1Indiana University
2University of Iowa
zhc@indiana.edu
Detecting the association between a set of variants and a phenotype of interest is the first and important step in genetic and genomic studies. Although it attracted a large amount of attention in the scientific community and several related statistical approaches have been proposed in the literature, powerful and robust statistical tests are still highly desired and yet to be developed in this area. In this paper, we propose a powerful and robust association test, which combines information from each individual single nucleotide polymorphisms (SNPs) based on a set of independent burden tests. We compare the proposed approach with some popular ones through a comprehensive simulation study and real data application. Our results show that in general the new test is more powerful; the gain in detecting power can be substantial in many situations, compared to other methods.

Estimation of ROC Curve and AUC via Heterogeneous Box-Cox Transformations
Xiaoyu Cai1, Zhaohai Li2 and Aiyi Liu3
1U.S. Food and Drug Administration
2George Washington University
3National Institutes of Health
nkcai@nmai.gov
The receiver operating characteristic (ROC) curve is a popular tool for evaluating the performance of the discrimination ability of a diagnostic biomarker. Parametric and nonparametric methods exist in the literature for estimation of an ROC curve and its associated summary measures using data usually from a case-control study. Since the ROC curve remains unchanged under a monotone transformation, the biomarker data from both cases (diseased subjects) and controls (non-diseased subjects) are often transformed based on a common Box-Cox transformation (or other appropriate transformation) prior to the application of a parametric estimation method such as the commonly used bi-normal model. However, careful examination of the data often reveals that the biomarker values in the diseased and non-diseased population can only be normally approximated via different transformations. In this situation, existing estimation methods cannot be directly applied to the transformed data. On the other hand, using a common transformation may result in...
biased estimation of the ROC curve and its indexes. In this paper, we deal with the situation that biomarker data from both diseased and non-diseased population are normally distributed after being transformed with different Box-Cox transformations. Under this assumption, we show that existing methods based on a common Box-Cox transformation are invalid in that they possess considerable asymptotic biases. We move on to propose a method to estimate the underlying ROC curve and its area under the curve (AUC), and investigate its asymptotic performance compared to the nonparametric estimator that ignores any distributional assumptions. Along the way we derive asymptotic bias and variance for each AUC estimator under consideration. The method is exemplified with data from a case-control study concerning maternal folate metabolism biomarkers and infants born with neural tube defects.

Optimal two-stage design for estimating the Area under the receiver operating characteristic curve
Yougui Wu
University of South Florida
drodande@health.usf.edu

In this paper, we consider the design for evaluating the performance of an ordinal diagnostic test in classifying the disease status. Statistical methods are well developed for estimating the area under the receiver operating characteristic curve (AUC) when the gold standard is available for every unit in the sample, or in a two-phase study when the gold standard is ascertained only in the second phase in a sub-sample using a fixed sampling scheme. However, these methods do not attempt to optimize the sampling scheme to minimize the variance of AUC estimator. We derived the analytic variance formula for the AUC estimator and used it to obtain the optimal sampling design. The efficiency of the optimal sampling design is evaluated by simulation that compares the optimal sampling with simple random sampling and with proportional allocation. Results of the simulation study show the optimal sampling design achieve a substantial amount of variance reduction with an over-sample of subjects with low or high ordinal levels. The optimal sampling is applied to a real-world example to evaluate the performance of a questionnaire score in screening for childhood asthma.

A Class of Bivariate Semiparametric Control Charts with Application to Biostatistics
Elisavet Sofikitou
University at Buffalo
esofikit@buffalo.edu

The nonparametric joint monitoring is a relative new research area and not much work has been done, especially in the multivariate case. The existing literature is limited either on one-sided Shewhart-type charts or on control charts with memory. In this work, a general class of Shewhart-type, semiparametric control charts is introduced for monitoring the process mean and/or variance. The test statistic of the proposed class is not only based on the location of a single bivariate pair, but also on the location of the test sample observations when compared to/among the control limits. As a result, the charts are capable of detecting potential shifts in both location and variability. Indeed, in the case that one pair is used in the test statistic (which means that exactly one observation per characteristic should lie between the control limits), the charts can monitor only the mean. For the decision making and in order to make fair comparisons, one needs to study the characteristics of the control charts, such as Operating Characteristic Function (OCF), False Alarm Rate (FAR) and Average Run Length (ARL). To provide some formulae for the aforementioned quantities, one has to resort to the theory of bivariate order statistics and the theory of copulas. For this purpose, some new joint probability functions were introduced, based on fourfold binomial models, which are of interest from the multivariate distribution point of view and can be used, for example, to study multivariate control charts, complex reliability system, etc. The aim here is to exploit these novel distributions in order to express several complicated probabilities associated with the new schemes.

Session 53: Statistics in Biosciences Invited Session: Statistical Advancement in Bioscience Research

New methods for estimating follow-up rates in cohort studies
Xiaonan Xue
Albert Einstein College of Medicine
xiaonan.xue@einstein.yu.edu

The follow-up rate, a standard index of the completeness of follow-up, is important for assessing the validity of a cohort study. The commonly used “Percentage Method” is proportion of participants without loss to follow-up at all, therefore, does not take into account partial follow-up. Alternatively, the median follow-up time does not indicate the completeness of follow-up, and the reverse Kaplan-Meier (KM) based method and Clark’s Completeness Index (CCI) also have limitations. Here we propose a new definition for the follow-up rate, the Person-Time Follow-up Rate (PTFR), which is the observed person-time divided by total person-time assuming no dropouts. The PTFR cannot be calculated directly since the event times for dropouts are not observed. Therefore, two estimation methods are proposed: a formal person-time method (FPT) and a simplified person-time method (SPT). Simulations were conducted to assess the accuracy of the proposed and existing methods and show that the FPT has the highest accuracy overall. In most situations, the computationally simpler SPT and CCI methods are only slightly biased and can be used in tandem to obtain an accurate and tight interval for PTFR. However, the FPT is recommended when event rates and dropout rates are high.

Spatiotemporal Statistical Methods for Monitoring Glaucoma Progression Using Visual Field Data
Joshua Warren1, Samuel Berchuck2 and Jean-Claude Mwanza3
1Yale University
2Duke University
3University of North Carolina at Chapel Hill
joshua.warren@yale.edu

Diagnosing glaucoma progression early is critical for limiting irreversible vision loss. A common method for assessing glaucoma progression relies on a longitudinal series of visual fields (VF) acquired from a patient at regular intervals. VF data are characterized by a complex spatiotemporal correlation structure due to the data generating process and ocular anatomy. Thus, advanced statistical methods are needed to make clinical determinations regarding progression status and to predict future vision loss. We introduce multiple spatiotemporal methods (spatiotemporal boundary detection, spatially-varying change points) that permit prediction of the timing and spatial location of future vision loss and inform clinical decisions regarding disease progression. We apply the methods to data from the Vein Pulsation Study Trial in Glaucoma and the Lions Eye Institute trial registry and show how they can be used to improve estimation and prediction of multiple aspects of disease management in comparison to existing methods.

Applying CHW Method to 2-in-1 Design: Gain or Lose?
Xiaofei Bai1 and Qiqi Deng
2-in-1 design (Chen et al., 2017) is a design to adapt the development path based on interim results. It has gained increasing interest since introduced. The original 2-in-1 design used conventional statistic without adjustment to control type I error rate. This paper explains the difference between 2-in-1 design and traditional sample size estimation methods. In addition, we applied CHW method for type I error control on 2-in-1 strategy, and compared it with original method. It shows that CHW method in general leads to comparable power with conventional statistic proposed in the original 2-in-1 design. Under certain interim decision rules, CHW method is slightly more powerful than conventional test statistic. In particular, when the interim decision threshold is high enough that the rejection region of conventional test statistic becomes a subset of the rejection region of CHW, the CHW may be preferred. However the differences are generally small. On the other hand, conventional test statistic approach is easy to implement, and have advantages in interpretation. CHW method may have controversial result when the estimated interim treatment effect is big but the final treatment effect is relatively small or vice versa, while this can be avoided by using conventional test statistic in 2-in-1 design.

Session 54: Innovative statistical analysis and design in handling unconventional clinical data

Causal Effects in Clinical Endpoint Bioequivalence Studies in the Presence of Intercurrent Events
Yi Yue Lou1, Michael Jones2 and Wanjie Sun3
1 Vertex Pharmaceuticals, Inc.
2 University of Iowa
3 Center for Drug Evaluation and Research, FDA

In clinical endpoint bioequivalence (BE) studies, the primary analysis for assessing equivalence between a generic and an innovator product is based on the observed per-protocol (PP) population (usually completers and compliers). However, missing data and non-compliance are post-randomization intercurrent events and may introduce selection bias. Therefore, PP analysis is generally not causal. The FDA Missing Data Working Group recommended using ‘causal estimates of primary interest.’ In this paper, we propose a principal stratification causal framework and co-primary causal estimates to test equivalence, which was also recommended by the recently published ICH E9 (R1) addendum to address intercurrent events. We identify three conditions under which the current PP estimator is unbiased for one of the proposed co-primary causal estimates - the “Survivor Average Causal Effect” (SACE) estimand. Simulation shows that when these three conditions are not met, the PP estimator is biased, and may inflate Type I error and/or change power. We also propose a tipping point sensitivity analysis to evaluate the robustness of the current PP estimator in testing equivalence when the sensitivity parameters deviate from the three identified conditions, but stay within a clinically meaningful range. Our work is the first causal equivalence assessment in equivalence studies with intercurrent events.

When convention meets practicality: Pooled analysis testing under the two-study paradigm
Dong Xi, Frank Bretz and Willi Maurer
Novartis

Standard considerations on multiplicity challenges arising within a single clinical study include comparing several doses of a new drug for more than one outcome variable. A different source of multiplicity arises when evidence is collected across multiple studies. For example, it is common practice to request two statistically significant confirmatory Phase III clinical studies (so-called two-study paradigm), which provides a stringent replicability standard in pharmaceutical drug development. Phase III studies are usually designed with a large sample size to provide sufficient exposure of the new drug in the target population. As a result, managing two large long-term studies may pose statistical, logistical and practical challenges. Motivated by a case study, we discuss in this presentation alternative test strategies that control the error rate across studies at an appropriate significance level while maintaining study-level familywise error rate conventionally. The proposed approaches are simple to communicate with clinical teams and other stakeholders.

Using Surrogate Endpoints in Adaptive Design with Delayed Treatment Effect
Qing Li and Jianchang Lin
Takeda Pharmaceuticals

In phase III oncology studies, delayed treatment effects have often been observed. These delayed treatment effects require a long-term approach to evaluate treatment effects. In addition, these phenomena bring more challenges to the interim analysis using survival endpoints. An improper interim analysis may falsely stop a promising compound due to the late separation of survival curves. In this scenario, short-term surrogate endpoints which are believed to be predictive of the primary long-term outcome can be extremely useful. For trials with delayed treatment effect, using surrogate endpoints in the interim analysis can help make more informative Go/No Go decision based on the interim analysis results and re-estimate sample sizes during the interim analysis to maintain a high probability of success. We propose using a surrogate endpoint (e.g. ORR) in the interim analysis to improve conditional power in designing adaptive sample size re-estimation trial with time-to-event endpoint (e.g. PFS). Through theoretic modeling and extensive simulations, our work will demonstrate the practical feasibility and benefits of using surrogate endpoints in adaptive trials with delayed treatment effect.

Seamless Phase 2/3 Oncology Trial Design with Flexible Sample Size Determination
Zhaoyang Teng1, Yuan Tian2, Yi Liu3 and Guohui Liu4
1 Takeda Pharmaceuticals
2 North Carolina State University
3 Nektar Therapeutics
4 Takeda Pharmaceuticals

Conventional seamless phase 2/3 design with fixed sample size determination (SSD) has been increasingly used in oncology drug development by significantly shortening the development timeline, minimizing sample size as well as early Go/No-Go decision making and treatment selection. However, the design wouldn’t be robust when the treatment effect assumption is not accurate when only limited efficacy data are available at study design stage. We propose an innovative seamless phase 2/3 study design with flexible SSD for oncology trials, in which the trial is designed under a range of treatment effect instead of one single assumption and the sample size for phase 3 portion is not predetermined at design stage, but dynamically determined based on observed treatment effect at phase 2 portion. The practical sample size determination rule for phase
3 portion is well discussed. The proposed design can determine sample size wisely leading to the reduced sample size or/and improved power compared to conventional seamless phase 2/3 design with fixed SSD. This innovative study design can be implemented to programs with aggressive development strategy but with limited data available at design stage while improve the probability of success and faster delivering the efficacious treatment to patients.

**Session 55: Recent Advances in Statistical Normalization Methods for Sequencing Data**

**Copy number analysis of circulating cell-free DNA**

*Venkatraman Seshan, Nicholas Socci, Dana Tsui and Julie Yang*

Memorial Sloan Kettering Cancer Center

Venkatraman.Seshan@mskcc.org

Abstract: Cell-free plasma DNA (cfDNA) is defined as fragments of DNA present in the extracellular fluid. These DNA fragments are mostly derived from cells that underwent apoptosis. In a subject with cancer, cfDNA are derived from both normal and tumor cells and thus it will reflect somatic changes present in the tumor DNA. In this talk we will present a method to estimate tumor fraction from shallow whole genome sequencing and one to estimate allele specific copy numbers from targeted deep sequencing. We will demonstrate the statistical issues and computational challenges faced with data from data from samples that have undergone these assays.

**Super-delta: a robust method that combines data normalization and differential expression analysis**

Yuhang Liu1, Jinfeng Zhang1 and Xing Qiu2

1Florida State University
2University of Rochester

xinqiu@urmc.rochester.edu

In genomic analysis, normalizations are pre-processing methods designed to remove technical errors (laboratory effects, batch effects, etc) before gene differential expression (DE) analysis. This variance reduction is typically done by forcing a pre-defined distributional property across genes (such as per-sample mean/median/quantile) to be constant across samples. Since true differentially expressed genes (DEGs) can have meaningful different distributional properties among the samples; normalization across all genes inevitably introduces an initial bias that may affect the subsequent DE analyses. We propose a new method called "super-delta", which is a local normalization procedure that minimizes the bias introduced by true differential expressions, and can be fully integrated into the subsequent DE analysis. In both simulation and real data analysis, we showed that super-delta achieved better type I error / statistical power trade off than competing methods. In addition, we proved the asymptotic unbiasedness of super-delta method via large sample theory.

**Statistical assessment of depth normalization methods for microRNA sequencing**

Li-Xuan Qin

Memorial Sloan Kettering Cancer Center

qin1@mskcc.org

Quality data is the foundational cornerstone for reliable scientific findings in evidence-based medical research. It is widely accepted that a crucial step to derive high-quality genomics data is to identify data artifacts caused by systematic differences in the processing of specimens and to remove these artifacts by data normalization. One major and unique aspect of RNA sequencing data normalization is the 'depth of coverage', defined as the average number of reads for a molecule being sequenced. Statistical methods for depth normalization have been recently developed, including both simple re-scaling-based methods and regression-based methods. Many of these normalization methods rely on the presupposition that variations in the assumed scaling factor or in the projection of the assumed regression function are solely due to data artifacts and should be removed. MicroRNAs are a unique class of small RNAs regulating gene expression and closely linked to carcinogenesis. They are low-complexity molecules (that is, a small number of molecules expressed dominantly) that tend to be expressed in a tissue-specific manner, especially in heterogeneous samples such as tumors. As a result, the assumption of depth normalization methods may not hold for microRNA sequencing. We performed a study to assess the performance of existing depth normalization methods on identifying disease relevant microRNAs using both a pair of datasets on the same set of tumor samples and data simulated from the paired datasets under various scenarios of differential expression. In this talk we will report our findings from this study.

**SCnorm: A quantile-regression based approach for normalization of single-cell RNA-seq data**

Rhonda Bacher

University of Florida

rbacher@ufl.edu

Single cell RNA-sequencing (scRNA-seq) is a promising tool that facilitates study of the transcriptome at the resolution of a single cell. However, along with the many advantages of scRNA-seq come technical artifacts not observed in bulk RNA-seq studies. The normalization methods traditionally used in bulk RNA-seq were not designed to accommodate these features and, consequently, applying them to the single-cell setting results in artifacts that bias downstream analyses. To address this, we developed SCnorm to enable efficient and accurate scRNA-seq normalization. Simulation and case study results demonstrate that SCnorm provides for increased accuracy in fold-change estimation as well as improvements in downstream inference.

**Normalization approaches for gene expression studies completed with high-throughput sequencing**

Brooke Fridley1, Farnoosh Aghababazadeh1 and Qian Li2

1Moffitt Cancer Center
2University of South Florida

Brooke.Fridley@moffitt.org

Normalization of RNA-Seq data has proven essential to ensure accurate inferences and replication of findings. Various normalization methods have been proposed for various technical artifacts that can be present in high-throughput sequencing transcriptomic studies. In this presentation, I will present some of the normalization methods used in RNA-Seq studies with application to data from The Cancer Genome Atlas (TCGA) cervical cancer study. Simulation study results will also be presented to compare the performance of the across sample normalization methods in estimating technical artifacts. Lastly, I will discuss the impact of normalization resulting in the reduction in degrees of freedom and the impact on downstream differential expression analysis.
Session 56: Biomarkers in Cancer Clinical Trials

Bayesian adaptive enrichment design for continuous biomarkers

Lindsay Renfro and Yusha Liu
1University of Southern California
2Rice University

Randomized clinical trials are the cornerstone of evidence-based medicine and the gold standard for establishing causal relationships between new treatments and improved patient outcomes. However, as diseases like cancer are increasingly understood on a molecular level, clinical trials are being prospectively designed to reveal or validate subpopulations in which the experimental therapy is more or less effective. Such "biomarker-driven" designs, particularly those of the "adaptive enrichment" variety, which initially enroll an unselected population and then potentially adapt to enroll only "marker-positive" patients based on the results of interim analyses, is increasingly popular. However, existing approaches often assume that the driving biomarker of interest is dichotomous in nature ("positive" or "negative" for each patient), when in reality, the underlying biomarker is continuous and may additionally have a non-linear relationship with outcome or treatment effect. We propose an adaptive enrichment design for continuous biomarkers that takes uncertainty regarding the shape and strength of these relationships into account, where following 1 or more interim analyses, decisions to continue enrollment in all patients, enrich to a subgroup, or terminate for efficacy or futility are considered. Using simulations and patient-level data from a completed clinical trial, we derive the operating characteristics of our design framework and evaluate its practical performance compared to a traditional adaptive enrichment approach that assumes up-front dichotomization of the marker of interest.

On the Design of Stratified Biomarker Trials With Measurement Error

Susan Halabi, Chen-Yen Lin and Aiyi Liu
1Duke University
2Eli Lilly and Company
3National Institutes of Health

The stratified biomarker design has become a popular design due to its ability to address various questions. The stratified biomarker design has also been used to test for a treatment-biomarker interaction in predicting an outcome. Many biomarkers, however, are tissue-based, and hence are heterogeneous. Thus, biomarker levels are measured with error and would have an adverse impact on the power of a stratified biomarker clinical trial. We investigate analytically and numerically the impact of biomarker misclassification on the coverage of the confidence intervals and the power for testing biomarker-treatment arm interaction. We propose sample size formulas for both binary and time-to-event endpoints subject to censoring, and apply the proposed sample size formulae to the design of a renal cancer trial.

Two Stage Adaptive Design for Prognostic Biomarker Signature with Survival Outcome

Biyue Dai and Mei-Yin Polley
1University of Iowa
2Mayo Clinic

Cancer biomarker discoveries typically involve utilizing patient specimens. The process of collecting and preserving high quality biospecimens is resource intensive, and hence there is often strong desire to preserve these precious materials to be used for studies that are most likely to yield useful information. To meet this practical demand, previously we proposed a two-stage cutoff design suitable for binary endpoints which attempts to terminate the biomarker study early in a futility interim if the performance of the model is deemed unsatisfactory, thereby allowing researchers to preserve the remaining specimens. In this work, we extend the novel two-stage design framework to accommodate time-to-event endpoints which are most common in oncology studies. Specifically, the first stage of the procedure involves testing whether the measure of discrimination for survival models (C-index) exceeds a pre-specified threshold. We describe the computation of cross-validated C-index and evaluation of the statistical significance using re-sampling techniques. An independent model validation is carried out at the second stage. Simulation studies are conducted to evaluate the operating characteristics of the proposed design. We show that under the null hypothesis (when the biomarker signature has weak prognostic value), our proposed design maintains the type I error at the nominal level, has high probabilities of terminating the study early, and results in sizable savings in biospecimens. Under the alternative hypothesis, we note that power of the design is a function of the true event rate, the total sample size, and the targeted degree of improvement in the discriminant measure. We apply the method to design of a prognostic biomarker study in patients with triple negative breast cancer. Some practical aspects of the proposed method are also discussed.

Utilizing Integrated analysis of RNA and DNA from phase III oncology trials to assess response

Terry Hyslop, Chuck Perou, Katherine Hoadley and Maki Tanioka
1Duke University
2University of North Carolina at Chapel Hill

Background: Response to a complex drug regimen in oncology is affected by multiple features of the tumor and its microenvironment. Developing a predictive algorithm is key to optimizing targeted therapies. Methods: We analyzed 137 pre-treatment tumors with mRNA-seq and DNA exome sequencing from CALGB 40601, a neoadjuvant phase III trial of paclitaxel plus trastuzumab with or without lapatinib in stage II-III HER2-positive breast cancer. We adopted an Elastic Net regularized regression approach that controls for covarying features within high-dimensional data. First, we applied 517 known gene expression signatures to develop an Elastic Net model to predict pathologic complete response (pCR), which we validated on 143 samples from 4 independent trials. Next, we performed integrative analyses incorporating clinicopathologic information with somatic mutation status, DNA copy number alterations (CNAs) and gene signatures. Results: The Elastic Net model using only gene signatures predicted pCR in the validation sets (AUC = 0.76). Integrative analyses showed that models containing gene signatures, clinical features, and DNA information were better pCR predictors than models containing a single data type. Frequently selected variables from the multi-platform models included amplifications of chromosome 6p, TP53 mutation, HER2-enriched subtype and immune signatures. Variables predicting resistance included Luminal/ER+ features. Conclusions: Models using RNA only, as well as integrated RNA and DNA models, can predict pCR with improved accuracy over clinical variables. Somatic DNA alterations (mutation, CNAs), tumor molecular subtype (HER2E, Luminal), and the microenvironment (immune cells) were independent predictors of response to trastuzumab and paclitaxel-
based regimens. This highlights the complexity of predicting response in HER2-positive breast cancer. Conclusions: Elastic Net modeling can be optimized to predict drug response. Integrating trial design variables, clinical and pathologic features, somatic mutation status, DNA copy number alterations and gene signatures provides the most informative models, with strong predictive value that can be validated in additional trials.

Session 57: Recent Advances in Precision Medicine Research

Constructing personalized decision algorithm for mHealth applications

Xinyu Hu, Min Qian, Bin Cheng and Ken Cheung
Columbia University

Mental illnesses affect tens of millions of people each year. However, only half of those in need actually receive treatment. This is partly due to the substantial barriers associated with accessing office-based mental health care. As such, there are great needs for providing those who are in need of help with access to efficacious therapies. The use of mobile applications can fill the gap by delivering personalized treatments to patients who will otherwise not have access to the traditional treatments. In this work, we proposed a new analytical framework to develop personalized mobile decision algorithms to optimize immediate goals, such as response to a pushed notification or reminder. The method is evaluated using simulation studies and illustrated using data from a recent mobile health study.

Comparative Intervention Scoring for Heterogeneity of Long-term Health System Intervention Effects

Jared Huling1, Menggang Yu2 and Maureen Smith2
1Ohio State University
2University of Wisconsin-Madison

With the growing cost of health care in the United States, the need to improve efficiency and efficacy of the delivery of care has become increasingly urgent. To this end, there have been widespread efforts to design and implement interventions which coordinate the typically fragmented care of complex patients, yet the effectiveness of such interventions in practice has been mixed. A common thread among successful care coordination interventions is the targeting of patients likely to benefit for enrollment, however, there is little guidance toward effectively doing so. In this work we seek to fill this gap by introducing a procedure to estimate personalized scores which characterize differential benefit of long-term health system interventions. As patients tend to respond differently over time, our approach allows the differential effects of an intervention to vary with time and encourage these effects to be more similar for closer time points. We utilize our approach to construct personalized enrollment decision rules for a complex case management intervention in a large health system and demonstrate that the enrollment decision rules result in improvement in health outcomes and care costs.

Optimal dynamic treatment regimes using decision lists

Yichi Zhang1, Eric Laber2, Bath Tsai2 and Marie Davidian2
1University of Rhode Island
2North Carolina State University

A dynamic treatment regime (DTR) formalizes precision medicine as a series of functions over decision points. At each decision point, it takes the available information of a patient as input and outputs a recommended treatment for that patient. A high-quality DTR tailors treatment decisions to individual patient as illness evolves, and thus improves patient outcomes while reducing cost and treatment burden. To facilitate meaningful information exchange during the development of DTRs, it is important that the estimated DTR be interpretable in a subject-matter context. We propose a simple, yet flexible class of DTRs whose members are representable as a short list of if-then statements. DTRs in this class are immediately interpretable and are therefore appealing choices for broad applications in practice. We develop a nonparametric Q-learning procedure to estimate the optimal DTR within this class. We establish its consistency and rate of convergence. We demonstrate the performance of the proposed method using simulations and a clinical dataset.

Improved doubly robust estimation in learning individualized treatment rules

Yinghao Pan1 and Yingqi Zhao2
1University of North Carolina at Charlotte
2Fred Hutchinson Cancer Research Center

Due to patient’s heterogeneous response to treatment, there is a growing interest in developing novel and efficient statistical methodologies in estimating individualized treatment rules (ITRs). The central idea is to recommend treatment according to patient characteristics, and the optimal ITR is the one that maximizes the expected clinical outcome if followed by the patient population. We propose an improved estimator of the optimal ITR that enjoys two key properties. First, it is doubly robust, meaning that the proposed estimator is consistent if either the propensity score or the outcome model is correct. Second, it achieves the smallest variance among its class of doubly robust estimators when the propensity score model is correctly specified, regardless of the specification of the outcome model. Simulation studies show that the estimated optimal ITR obtained from our method yields better clinical outcome than its main competitors. Data from Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study is analyzed as an illustrative example.

Session 58: Statistical Machine Learning in Analyzing Large-scale Multi-Dimensional Data

A Temporal Latent Factor Model for Product Sales Forecasting

Xuan Bi1, Gediminas Adomavicius1, William Li2 and Annie Qu3
1University of Minnesota
2Shanghai Jiao Tong University
3University of Illinois at Urbana-Champaign

With the advent of the big-data era, personalization is becoming increasingly important for electronic commerce, as they assist users in finding customized content, products, and services among the ever-increasing set of available alternatives. One of the most effective predictive methodologies to achieve such personalized recommendation is the latent factor model, whose applications have focused largely on modeling individual, subjective, taste-driven consumer preferences in domains of “experience” goods (e.g., movies, music, or books). In this article, we apply the latent factor model to the domain of sales forecasting, which achieves individualized sales forecasting for brick-and-mortar stores and product distributors. The major contribution of our work is that we incorporate local sales competition into the model, which formulates local market demand for different product categories, in addition to improving prediction.
accuracy. Technically, the proposed method applies a tensor factorization approach for time-aware, simultaneous modeling of past and current sales across multiple products and multiple stores, utilizes an additional set of negatively-correlated region- or product-specific latent factors to formulate sales competition, and conducts a seasonal time-series model to extend prediction results to future time points. The advantages of the proposed method are demonstrated by comparing its performance to a number of techniques from prior literature on a large dataset of sales transactions from more than 2,000 grocery stores across 47 U.S. markets, in terms of both product sales forecasting accuracy and decision support (e.g., for store managers or product distributors) on new product introduction.

**Total Variation Smoothed Regression for Spatial and Temporal data**

*Ying Liu*, Bowei Yan, and Haochang Shou

1 University of Wisconsin
2 Jump Trading
3 University of Pennsylvania

The increasing availability of the high-throughput medical data has provided tremendous opportunity for researchers to search for biomarkers that link with pathology and help to improve diagnosis. Appropriate statistical methods that acknowledge the multidimensional structures as well as the correlations induced by temporal and spatial continuity are needed. Motivated by 1D physical activity data continuously assessed by wearable computing sensors and 3D structural magnetic resonance imaging (MRI) data, we proposed to use a total variation (TV) regularized image-on-scalar regression method that acknowledges the dependency while adjusting for confounding by covariates. The estimator is the solution of a penalized regression problem where the objective is the sum of square error plus a TV regularization on the predicted mean across all subjects. We developed an efficient and scalable algorithm via alternative direction of methods of multiplier (ADMM). The method has been applied to accelerometer data from NIMH family study of spectrum disorders and the gray matter voxel-based morphometry maps from the attention deficit/hyperactive deficit (ADHD) 200 consortium.

**Statistical Inference for 3D Rock Images using Persistent Homology**

*Samantha Morrison*, Jon Steingrimsson and Constantine Gatsonis

Brown University

samantha.morrison@brown.edu

Recently there has been large growth in the development of deep learning to analyze medical images, such as CT and MRI scans. These methods are used for detection, diagnosis, and prediction of patient outcomes in a broad range of diseases, including cancer. The prediction of time-to-event-outcomes, such as survival, requires methods for addressing the potential bias due to right censoring. Such methods have not received much attention in the context of deep learning. In this talk, we discuss methods for handling right censoring in the prediction of survival probabilities using deep learning. We use scans from digitized histology slides of brain tumor patients and implement several neural networks combined with censoring unbiased loss functions. In particular, we used the VGG16 neural network to extract features from the images and then trained fully connected neural networks with censoring unbiased loss functions (Buckley-James and Doubly Robust). We compared the resulting predictions to a Cox Proportional Hazards model using weighted Brier scores, and we found similar performance. These results, paired with the relative ease of implementing the methods, illustrates the potential for using these neural networks with censoring unbiased loss functions in a broad range of medical imaging studies.

**Session 59: Causal Inference with Complex Data Structures**

**Causal mediation analysis under partial compliance in randomized trials**

Fan Yang, Guanglei Hong

1 University of Colorado Denver
2 University of Chicago

fan.j.yang@ucdenver.edu

Causal mediation analysis plays an important role in program development to understand causal pathways through which an effect arises. To evaluate an indirect effect transmitted through a hypothesized mediator and a direct effect, causal mediation studies often take advantage of experimental designs. However, partial compliance has been common in experimental research. Given the post-treatment assignment nature of the treatment variable, the direct and indirect effects of the treatment assignment can not be identified without further assumptions beyond sequential ignorability. Under a deterministic principal stratification framework, Yamamoto (2014) proposed a local sequential ignorability assumption which identifies the direct and indirect effects of the treatment assignment among compliers. However, the deterministic framework is incompatible with key concepts in causal mediation analysis. To reflect the stochastic reality of program participation and the intermediate experiences, we propose assumptions to decompose the effect of treatment among the “strength-of-IV weighted population” (Small et al, 2017) under a stochastic framework. We apply our method to National Job Corps Study.

**Weighting-Based Sensitivity Analysis for Evaluating the Average Treatment Effect**

Guanglei Hong, Fan Yang and Xu Qin

1 University of Chicago
Abstracts

In quasi-experimental research or in experimental studies with non-random attrition, inference about causality often relies on the strong assumption that there is no hidden bias due to omitted confounders. Through a sensitivity analysis, the analyst attempts to determine whether a conclusion could be easily reversed by a plausible violation of this key assumption. A new approach to sensitivity analysis, on the basis of weighting, extends and supplements existing propensity score weighting methods for identifying the average treatment effect (ATE). In its essence, the discrepancy between a new weight that adjusts for the omitted confounders and an initial weight that omits them captures the role of the confounders. The effect size of a hidden bias is represented as a function of a small number of meaningful sensitivity parameters. This new approach conveniently assesses bias associated with one or more omitted confounders. The number of sensitivity parameters does not increase with the complexity of the data generation forms. Hence this approach reduces the reliance on functional form assumptions and removes constraints on measurement scales. We extend this innovative sensitivity analysis strategy to evaluations of ATE in the presence of (a) omitted posttreatment covariates associated with nonrandom attrition and (b) additional omitted time-varying confounding associated with noncompliance. Its broad utility is illustrated through a re-analysis of the well-known Project STAR data that evaluates the impact of class size reduction in elementary education.

Analysis of regression discontinuity designs using censored data

Youngjoo Cho1, Chen Hu2 and Debashis Ghosh3
1University of Wisconsin-Milwaukee
2Johns Hopkins University
3Colorado School of Public Health

In medical studies, the treatment assignment may be determined by a clinically important covariate that predicts patients’ risk of survival. There is a class of methods from the social science literature known as regression discontinuity (RD) designs that can be used to estimate the treatment effect in this situation. Under certain conditions, such an effect enjoys a causal interpretation. However, few authors have discussed the use of RD for censored data. In this paper, we show how to estimate causal effects under the regression discontinuity design for censored data. The proposed estimation procedure employs a class of censoring unbiased transformations that includes inverse probability weighting and a doubly robust transformation. Simulation studies demonstrate the utility of the proposed methodology.

Synthetic estimation for causal inference

Bing Han, Mathew Cefalu and Denis Agniel
RAND Corporation

A major difficulty for practitioners of Rubin Causal Model (RCM) is to choose from the large number of available estimators. Numerical and empirical studies showed that the conclusions across methods can be highly variable and that many distinct approaches have been recommended by different authors. To address this challenge, we propose a synthetic estimator based on the classic linear model averaging theory. The synthetic estimator is a convex combination of multiple candidate estimators with the goal of achieving an optimal mean squared error. We discuss the properties and computational details of the proposed synthetic estimator. We demonstrate by numerical examples that the synthetic estimator has a robust performance across various data generating strategies, while any single candidate estimator’s performance is usually volatile.

Session 60: Gaussian graphical models and sparsity

Sparse graphical modeling of longitudinal data
Yafei Zhang and Pang Du
University of Utah

Gaussian graphical models represent conditional dependence relationships among the associated variables. In the literature, a lot of research developments have been made for high dimensional Gaussian graphical models based on i.i.d observations. In some practical problems, the underlying graphical structure can undergo “some changes”, and the methods designed for i.i.d cases are no longer applicable. In this talk, we investigate the problem of simultaneous change point identification and structure recovery in the context of high dimensional Gaussian graphical models with abrupt changes. To recover the graphical structure correctly, a data-driven thresholding procedure is introduced. We establish estimation consistency of the change point estimator, by allowing the number of nodes being much larger than the sample size. Furthermore it is shown that, in terms of structure recovery of Gaussian graphical models, the proposed procedure achieves model selection consistency and controls the number of false positives. The validity of our proposed method is justified by extensive numerical studies.

High-dimensional Q-learning for dynamic treatment regimes
Rui Song
North Carolina State University

An approximate EM algorithm for latent Gaussian graphical model
Chaowen Zheng1 and Yichao Wu2
1North Carolina State University
2University of Illinois at Chicago

This talk considers the latent Gaussian graphical model which extends the Gaussian graphical model to handle mixed data with both continuous and discrete variables by assuming that discrete variable is generated through discretizing a latent Gaussian variable. For this model, we propose an approximate EM algorithm to estimate the
parameters in the latent Gaussian model. Then the conditional dependence structure can be constructed by analyzing the sparsity pattern of the precision matrix of the latent Gaussian variables. We illustrate the superior performance of our proposed estimator through comprehensive numerical studies.

Session 61: Advanced methods in high dimensional and big data analysis

On post dimension reduction statistical inference

Bing Li, Kyongwon Kim, Lexin Li and Zhou Yu
Pennsylvania State University

The methodologies of sufficient dimension reduction have undergone extensive developments in the past three decades. However, there has been a lack of systematic and rigorous development of post dimension reduction inference, which has seriously hindered its applications. The current common practice is to treat the estimated sufficient predictors as the true predictors and use them as the starting point of the downstream statistical inference. However, this naive inference approach would grossly overestimate the confidence level of an interval, or the power of a test, leading to the distorted results. In this paper, we develop a general and comprehensive framework of post dimension reduction inference, which can accommodate any dimension reduction method and model building method, as long as their corresponding influence functions are available. Within this general framework, we derive the influence functions and present the explicit post reduction formulas for the combinations of numerous dimension reduction and model building methods. We then develop post reduction inference methods for both confidence interval and hypothesis testing. We investigate the finite-sample performance of our procedures by simulations and a real data analysis.

Copula-based semiparametric analysis for time series data with detection limits

Fuyuan Li, Yanlin Tang and Huixia Judy Wang
1 Pennsylvania State University
2 East China Normal University
3 George Washington University

The analysis of time series data with detection limits is challenging due to the high-dimensional integral involved in the likelihood. To account for the computational challenge, various methods have been developed but most of them rely on restrictive parametric distributional assumptions. We propose a semiparametric method for analyzing censored time series, where the temporal dependence is captured by parametric copula while the marginal distribution is estimated nonparametrically. Utilizing the properties of copula modeling, we develop a new copula-based sequential sampling algorithm, which provides a convenient way to calculate the censored likelihood. Even without full parametric distributional assumptions, the proposed method still allows us to easily compute the conditional quantities of the censored response at a future time point, and thus construct both point and interval predictions. We establish the asymptotic properties of the proposed pseudo maximum likelihood estimator, and demonstrate through simulation and the analysis of a water quality data that the proposed method is more flexible and it leads to more accurate predictions than Gaussian-based methods for non-normal data.

Big EHR Data Analysis and Modeling: Challenges and Opportunities

Hulin Wu
University of Texas Health Science Center at Houston

The real world EHR and health care Big Data may bring a revolutionary thinking on how to evaluate therapeutic treatments in a real world setting. Big EHR data may also allow us to identify specific patient populations for a specific treatment so that the concept of personalized treatment can be implemented and deployed directly on the EHR system. However, it is quite challenging to use the real world data in treatment assessment and disease predictions due to various reasons. In this talk, I will share our experiences on EHR and health care Big Data research. In particular, I will discuss the basic infrastructure and multi-disciplinary team that need to be established. Then I will demonstrate how to identify meaningful clinical questions and develop the analytic pipelines to address a class of clinical questions based on Big Health Care Data. Examples from disease-disease interaction network modeling, heart failure prediction, and vasopressor treatment evaluation for subarachnoid hemorrhage (SAH) patients based EHR data will be used to illustrate the novel concepts, challenges and opportunities to use the real world data for research to address important scientific questions.

Session 62: Recent Advances in Statistical Methodology for Analyzing Human Metagenomics Data

Powerful adaptive microbiome differential abundance analysis

Xiang Zhan
Pennsylvania State University College of Medicine

Differential abundance analysis is a crucial task in many microbiome studies, where the central goal is to identify microbiome taxa associated with certain biological or clinical conditions. There are two different modes of microbiome differential abundance analysis: the individual-based univariate differential abundance analysis and the group-based multivariate differential abundance analysis. The univariate analysis identifies differentially abundant microbiome taxa under certain statistical error measurements such as false discovery rate, which is typically complicated by the high-dimensionality of taxa and complex correlation structure among taxa. The multivariate analysis evaluates the overall shift in the abundance of microbiome composition between two conditions, which provides useful preliminary information for the necessity of follow-up univariate analysis. In this paper, we present a novel Adaptive multivariate two-sample test for Microbiome Differential Analysis (AMDA) to examine whether the composition of a group of taxa are different between two conditions. Despite being a group-based multivariate test that does not target at identifying individual differentially abundant microbiome taxa, the intermediate variable selection procedure in AMDA can provide useful information regarding the importance of individual taxa in the taxa-group. Our simulation studies and real data applications demonstrated that the AMDA test was often more powerful than several competing methods while preserving the correct type I error rate. A free implementation of our AMDA method in R software is available upon request.

HMP2Data: Integrative Human Microbiome Data R Bioconductor package and analysis workflow

Ekaterina Smirnova, Ni Zhao, Mikhail Dozmorov and Levi
The integrative Human Microbiome project (iHMP) generated longitudinal datasets from three different cohorts to study the association between microbiome and (1) pregnancy and preterm birth; (2) inflammatory bowel disease; and (3) type 2 diabetes. However, working with these data is daunting due to complex processing steps to (1) access, import and merge different components in formats suitable for ecological and statistical analysis; and (2) visualize and combine the data to analyze the longitudinal, multi-omics and multi-body site host-microbiota interactions. We present the community-resource package HMP2Data that allows researchers to easily access the iHMP data deposited at the coordinating center (DACC). Each cohort data was harmonized using MultiAssayExperiment and phylseq packages to allow easy data management and analysis. We concentrate on the vaginal microbiome data from the pregnancy and preterm birth study and discuss recent analyses of similarities across cytokines and 16S data using co-inertia techniques.

A framework for multi-view analysis of microbiome data

Timothy Randolph1, Yue Wang2, Ali Shojaie3 and Jing Ma4

1Fred Hutchinson Cancer Research Center
2University of Washington
3CUNY Graduate School of Public Health and Health Policy
4Virginia Commonwealth University

Ekaterina Smirnova

The analysis of microbiome data is often carried out from multiple perspectives. For example, data typically includes measurements from hundreds of taxa in each sample, but these measures may take multiple forms: taxon abundance, taxon presence/absence, relatedness among taxa, taxon gene counts, and/or metabolite production (or potential thereof). These multiple “views” of each sample don’t fit neatly into standard statistical analyses or dimension-reduction methods. This talk presents a framework of incorporating more than one view of these data into tools such as a principal coordinates (PCoA) plots and penalized regression models. A motivating example is how to construct a biplot that includes the contribution of taxon abundances into a PCoA plot based on unweighted Unifrac dissimilarities.

A robust distance-based kernel association test for correlated microbiome data

Hyunwook Koh1, Yutong Li2, Xiang Zhan3, Jun Chen4 and Ni Zhao1

1Johns Hopkins University
2University of Pennsylvania
3Pennsylvania State University
4Mayo Clinic

Researchers have increasingly employed family-based or longitudinal study designs to survey the roles of the human microbiota on diverse host traits of interest (e.g., health/disease status, medical intervention, behavioral/environmental factor). Such study designs are useful to properly control for potential confounders or the sensitive changes in microbial composition and host traits. However, downstream data analysis is challenging because the measurements within clusters (e.g., families, subjects including repeated measures) tend to be correlated so that statistical methods based on the independence assumption cannot be used. For the correlated microbiome studies, a distance-based kernel association test based on the linear mixed model, namely, correlated sequence kernel association test (cSKAT), has recently been introduced. cSKAT models the microbial community using an ecological distance (e.g., Jaccard/Bray-Curtis dissimilarity, unique fraction distance), and then tests its association with a host trait. Similar to prior distance-based kernel association tests (e.g., microbiome regression-based kernel association test (MiRKAT)), the use of ecological distances gives a high power to cSKAT. However, cSKAT is limited to handling Gaussian traits (e.g., body mass index (BMI)) and a single chosen distance measure at a time. The power of cSKAT differs a lot by which distance measure is used, yet it is also highly challenging to choose an optimal distance measure to use because of the unknown nature of the true association. Here, we introduce a distance-based kernel association test based on the generalized linear mixed model (GLMM), namely, GLMM-MiRKAT, to handle diverse types of traits, such as Gaussian (e.g., BMI), Binomial (e.g., disease status, treatment/placebo) or Poisson (e.g., number of tumors/treatments) traits. We further propose a data-driven adaptive test of GLMM-MiRKAT, namely, aGLMM-MiRKAT, so as to avoid the need to choose the optimal distance measure. Our extensive simulations demonstrate that aGLMM-MiRKAT is robustly powerful while correctly controlling type I error rates. We apply aGLMM-MiRKAT to real familial and longitudinal microbiome data, where we discover significant disparity in microbial community composition by BMI status and the frequency of antibiotic use. aGLMM-MiRKAT is a useful analytical tool with its broad applicability to diverse types of traits, robust power and valid statistical inference.
The presentation will discuss some statistical methods in development for borrowing from adult or other populations to make inferences about a pediatric population. We will take a regulatory and Bayesian perspective, and discuss opportunities and challenges with implementing "extrapolation" as described in the 2016 FDA Guidance "Leveraging Existing Clinical Data for Extrapolation to Pediatric Uses of Medical Devices".

Design innovation in pediatric drug development strategy: what can be learned and what can be better

Sarah Stanley and Margaret Gamalo-Siebers
Eli Lilly and Company
sarah.stanley@lilly.com

With the rise of regulatory requirements within the pediatric drug development space in the last decade, the attention of academic and industry researchers alike has centered on the goal of developing efficient pediatric trial designs that balance the need for information with the ethics of enrolling children as a vulnerable population. Innovative techniques such as extrapolation have come into focus as clinicians seek to reduce the need for pediatric exposure by leveraging prior information and developing analysis models that take this existing information into account. The drive for innovation and efficiency, has resulted in numerous examples of analytical and design strategies that have greatly impacted pediatric drug development. This presentation will explore what can be learned from these examples and how knowledge of past experiences can be utilized to improve design efficiency amidst a constantly evolving clinical and regulatory space. The trial design of a recently approved pediatric drug is investigated as a case example along with an exploration of regulatory space. The trial design of a recently approved pediatric drug is investigated as a case example along with an exploration of regulatory space. The trial design of a recently approved pediatric drug is investigated as a case example along with an exploration of regulatory space. The trial design of a recently approved pediatric drug is investigated as a case example along with an exploration of regulatory space. The trial design of a recently approved pediatric drug is investigated as a case example along with an exploration of regulatory space. The trial design of a recently approved pediatric drug is investigated as a case example along with an exploration of regulatory space. The trial design of a recently approved pediatric drug is investigated as a case example along with an exploration of regulatory space. The trial design of a recently approved pediatric drug is investigated as a case example along with an exploration of regulatory space. The trial design of a recently approved pediatric drug is investigated as a case example along with an exploration of regulatory space.

Session 64: Recent advances in functional data analysis

A Multiplier Bootstrap Approach to One-way ANOVA for Functional Data

Zhenhua Liu, Miles Lopes and Hans-Georg Müller
University of California, Davis
linzh@ucdavis.edu

We propose a new approach to the problem of one-way ANOVA for functional data based on our recent development of multiplier bootstrap for high-dimensional data with weak variance decay. We show that, the test not only admits the root-n consistency, but also has a size of a nearly parametric rate. Numerically, it is comparable to existing tests in general situations, while exhibits a clear advantage when the problem is hard in the sense that the functional data are rough and/or the signals are weak.

Estimation and Inference for Functional Linear Regression Models with Varying Regression Coefficient

Guangqun Cao1, Shuoyang Wang1 and Li Wang2
1 Auburn University
2 Iowa State University
gzc0009@auburn.edu

In this work, we present a class of functional linear regression models of a functional response on one or multiple functional predictors and scalar predictors. In particular, the approach can accommodate densely or sparsely sampled functional responses as well as multiple scalar and functional predictors. It also allows for the integration of continuous or categorical covariates. Tensor product B-spline basis is proposed for the estimation of the bivariate coefficient functions. We show that our estimators hold asymptotic consistency and normality. Several numerical examples demonstrate superior performance of the proposed methods against two existing approaches.

Nonlinear and additive principal component analysis for functional data

Jun Song1 and Bing Li2
1 University of North Carolina at Charlotte
2 Pennsylvania State University
songjun0525@gmail.com

In this talk, I will discuss a nonlinear and additive version of principal component analysis for vector-valued random functions. This is a generalization of functional principal component analysis that allows the relations among the random functions involved to be nonlinear. The method is constructed via two additively nested Hilbert spaces of functions, in which the first space characterizes the functional nature of the data, and the second space captures the nonlinear dependence. In the meantime, additivity is imposed so that we can avoid high-dimensional kernels in the functional space, which causes the curse of dimensionality. Simulation results show that the new method outperforms functional principal component analysis when the relations among random functions are nonlinear. Application to online handwritten digit data will be presented.

B-scaling: a novel nonparametric data fusion method

Yiwen Liu1, Xiaoxiao Sun2, Wenxuan Zhong3 and Bing Li3
1 University of Arizona
2 University of Georgia
3 Pennsylvania State University
xiasun@email.arizona.edu

With the rapid advancement in science and technology, the proliferation of large-scale data with various types and formats has been growing at an unprecedented rate. Each of these distinct data types provides a different but complementary view of the corresponding scientific problem, which makes systematic integration an essential component in many fields of science. Despite the need for powerful and advanced integrative analysis strategies, it still lacks the coherent statistical framework for data integration. In this talk, I will introduce a novel and unified framework for data integration. In the proposed framework, we observe multiple measurements that measure the same underlying quantity through some linear or nonlinear ways. We seek to find a fused measurement, which is a scalar representation of those observed measurements and highly correlated with the underlying quantity. We also established the asymptotic property of the fused measurement and integrated multiple histone modifications and DNA methylation levels to jointly characterize epigenetic activeness.

Session 65: Making sense of data for business: the role of statistics

Factor models for matrix-valued high-dimensional time series

Xialu Liu
San Diego State University
xialu.liu@sdsu.edu

In finance, economics and many other fields, observations in a matrix form are often observed over time. For example, many economic indicators are obtained in different countries over time. Various financial characteristics of many companies are reported over time. Although it is natural to turn a matrix observation into a long vector then use standard vector time series models or factor analysis, it is often the case that the columns and rows of a matrix represent different sets of information that are closely interrelated in...
a very structural way. We propose a novel factor model that maintains and utilizes the matrix structure to achieve greater dimensional reduction as well as finding clearer and more interpretable factor structures. Estimation procedure and its theoretical properties are investigated and demonstrated with simulated and real examples.

**Integrative interaction analysis using threshold gradient directed regularization**

*Yang Li*, Rong Li<sup>1</sup>, Yichen Qin<sup>2</sup>, Mengyun Wu<sup>3</sup> and Shuangge Ma<sup>4</sup>

<sup>1</sup>Renmin University of China  
<sup>2</sup>University of Cincinnati  
<sup>3</sup>Shanghai University of Finance and Economics  
<sup>4</sup>Yale University

For many complex business and industry problems, high-dimensional data collection and modeling have been conducted. It has been shown that interactions may have important implications beyond the main effects. The number of unknown parameters in an interaction analysis can be larger or much larger than the sample size. As such, results generated from analyzing a single data set are often unsatisfactory. Integrative analysis, which jointly analyzes the raw data from multiple independent studies, has been conducted in a series of recent studies and shown to outperform single-data set analysis, meta-analysis, and other multi-data set analyses. In this study, our goal is to conduct integrative analysis in interaction analysis. For regularized estimation and selection of important interactions (and main effects), we apply a threshold gradient directed regularization approach. Advancing from the existing studies, the threshold gradient directed regularization approach is modified to respect the “main effects, interactions” hierarchy. The proposed approach has an intuitive formulation and is computationally simple and broadly applicable. Simulations and the analyses of financial early warning system data and news-APP recommendation behavior data demonstrate its satisfactory practical performance.

**Protecting Privacy of Household Panel Data**

*Shaobo Li<sup>1</sup>, Matthew Schneider<sup>2</sup>, Yan Yu<sup>3</sup> and Sachin Gupta<sup>4</sup>*

<sup>1</sup>University of Kansas  
<sup>2</sup>Drexel University  
<sup>3</sup>University of Cincinnati  
<sup>4</sup>Cornell University

We investigate how vulnerable the widely used consumer/household panel data is to intruders who intend to re-identify individual customers and learn their private information through linking to external databases. With the evidence of low privacy-preserving being shown, we establish a framework to protect household panel data without reducing the commercial value. That is, retailers and manufacturers are able to use the household panel data to perform marketing research such as modeling brand choice, while they are not able to re-identify individuals by using external dataset. The proposed data protection method is based on a Bayesian hierarchical multinomial logit model that synthesizes the brand choice response at various protection level while preserving the household heterogeneity, which is crucial in choice modeling and other marketing objectives. We demonstrate our framework using IRI household panel data, where we use the yogurt category for assessing the data utility with a brand choice model.

**Model Confidence Bounds for Model Selection**

*Yichen Qin*  
University of Cincinnati

In this article, we introduce the concept of model confidence bounds (MCB) for variable selection in the context of nested models. Similarly to the endpoints in the familiar confidence interval for parameter estimation, the MCB identifies two nested models (upper and lower confidence bound models) containing the true model at a given level of confidence. Instead of trusting a single selected model obtained from a given model selection method, the MCB proposes a group of nested models as candidates and the MCB’s width and composition enable the practitioner to assess the overall model selection uncertainty. A new graphical tool - the model uncertainty curve (MUC) - is introduced to visualize the variability of model selection and to compare different model selection procedures. The MCB methodology is implemented by a fast bootstrap algorithm that is shown to yield the correct asymptotic coverage under rather general conditions. Our Monte Carlo simulations and real data examples confirm the validity and illustrate the advantages of the proposed method.

**Session 66: Meet Both the Computational and Statistical Challenges in Learning**

**Parameterized matrix factorization with missing data via non-convex optimization**

*Xiaodong Li*

University of California, Davis

We introduce the concept of model confidence bounds (MCB) for variable selection in the context of nested models. Similarly to the endpoints in the familiar confidence interval for parameter estimation, the MCB identifies two nested models (upper and lower confidence bound models) containing the true model at a given level of confidence. Instead of trusting a single selected model obtained from a given model selection method, the MCB proposes a group of nested models as candidates and the MCB’s width and composition enable the practitioner to assess the overall model selection uncertainty. A new graphical tool - the model uncertainty curve (MUC) - is introduced to visualize the variability of model selection and to compare different model selection procedures. The MCB methodology is implemented by a fast bootstrap algorithm that is shown to yield the correct asymptotic coverage under rather general conditions. Our Monte Carlo simulations and real data examples confirm the validity and illustrate the advantages of the proposed method.

**High-dimensional Gaussian graphical model on network-linked data**

*Tianxi Li<sup>1</sup>, Cheng Qian<sup>2</sup>, Elizaveta Levina<sup>3</sup> and Ji Zhu<sup>3</sup>*

<sup>1</sup>University of Virginia  
<sup>2</sup>Southeastern University  
<sup>3</sup>University of Michigan

Graphical models are commonly used to represent conditional independence relationships between variables, and estimating them from high-dimensional data has been an active research area. However, almost all existing methods rely on the assumption that the observations share the same mean, and that they are independent. At the same time, datasets with observations connected by a network are becoming increasingly common, and tend to violate both these assumptions. In the paper, we develop a Gaussian graphical model for settings where the observations are connected by a network and...
have potentially different mean vectors, varying smoothly over the network. We propose an efficient estimation method for this model and demonstrate its effectiveness in both simulated and real data, obtaining meaningful interpretable results on a statistician's coauthorship network. We also prove that our method estimates both the inverse covariance matrix and the corresponding graph structure correctly under the assumption of network “cohesion”, which refers to the empirically observed phenomenon of network neighbors sharing similar traits.

Magic Cross-Validation for Support Vector Machines and Related Large Margin Classifiers

Boxiang Wang and Hui Zou

1University of Iowa
2University of Minnesota

boxiang-wang@uiowa.edu

In this paper, we study the use of cross-validation (CV) for the support vector machine (SVM) and related large margin classifiers. We address two wide-spread misconceptions on leave-one-out cross-validation (LOOCV). We argue that LOOCV and ten-fold or five-fold CV have comparable variance and the computation cost of LOOCV is not prohibitively more expensive than that of ten-fold and five-fold CV. We present a “magic” leave-one-out formula that leads to a new and very efficient algorithm named “magicsvm” for fitting and tuning the kernel SVM. By magicsvm, the computational cost of LOOCV is of the same order of magnitude as that of a single SVM on the training data, and magicsvm is shown to be much faster than the state-of-the-art SVM solvers based on extensive simulations and benchmark examples. We are interested in estimating the post-selection generalization error of the kernel classifier and propose an honest LOOCV estimator that is nearly unbiased and has very competitive performance even when competing with the famous 0.632+ estimator.

High dimensional independence testing with maxima of rank correlations

Mathias Drton, Fang Han and Hongjian Shi

1University of Washington

faghan@uw.edu

Testing mutual independence for high dimensional observations is a fundamental statistical challenge. Popular tests based on linear and simple rank correlations are known to be incapable of detecting non-linear, non-monotone relationships, calling for methods that can account for such dependences. To address this challenge, we propose a family of tests that are constructed using maxima of pairwise rank correlations that permit consistent assessment of pairwise independence. Built upon a newly developed Cramér-type moderate deviation theorem for degenerate U-statistics, our results cover a variety of rank correlations including Hoeffding’s D, Blum–Kiefer–Rosenblatt’s R, and Bergsma–Dassios–Yanagimoto’s $\tau^*$. The proposed tests are distribution-free, implementable without the need for permutation, and are shown to be rate-optimal against sparse alternatives under the Gaussian copula model. As a by-product of the study, we reveal an identity between the aforementioned three rank correlation statistics, and hence make a step towards proving a conjecture of Bergsma and Dassios.

Session 67: Identification of biomarkers and disease mechanisms through proper biostatistics analysis

Characterizing the Assay Reproducibility of a Companion Diagnostics Test

Alan Chiang

Celgene

achiang@celgene.com

Companion diagnostics are assays or tests that are specifically developed for use as a companion to a particular drug. Helping identify the right patient population, companion diagnostics may be used to avoid adverse drug reactions by allowing doctors to identify patients who are at increased risk for serious side effects from certain medicines. Attempts to evaluate the reproducibility under ‘real world’ conditions are discussed in this talk. Multiple factors, including operators, instruments, assay reagent lots, and days play key roles in contributing both intra- and inter-laboratory reproducibility. Experimental design considerations are critical to successful assessment of assay performance.

Analysis of Flow Cytometry Data

Shuguang Huang

Stateward, LLC

shu44@gmail.com

Flow cytometry assay has become a primary means of assessing T cell responses in vaccine and immunotherapy product developments. Typically, the assay is used to determine (1) if a sample shows ‘positive’ response to a treatment or a stimulation, (2) if the assay readout is correlated with clinical endpoint. Flow cytometry assays have some unique challenges compared to other bioassays. For instance, technical replicates are rarely available, how can we design a study to evaluate the assay’s sensitivity (LOB, LOD, LOQ) and reproducibility? How can we decide whether a stimulation results in a positive response (compared to the unstimulated control) by using only one replicate? How can we develop a mathematical prediction model based on these highly correlated polyfunctions? In this presentation, the following issues will be discussed The statistical design for the analytical validation of cytometry assays The gating of “positive events” - the pros and cons of FlowJo and publicly available software The determination assay sensitivity - Limit of blanks (LOB), lower quantification limits (LOQ), and the detection limit (LOD) of the assay The identification of the “poly functions” that are associated with clinical endpoints (PK, PD, clinical outcome) The development of the predictive model that serves as a Biomarker (multivariate modeling)

Post-selection Inference for Regression Models with Linear Constraints

Jiarui Lu and Hongzhe Li

University of Pennsylvania

jiaruilu@pennmedicine.upenn.edu

Linear models with linear constraints naturally arise in many applications. Examples include regression analysis of microbiome compositional data, where linear models with a group of linear constraints have been developed to explore the association between taxa and outcome of interests. However, statistical inference for these models is not trivial due to both the high dimensionality and the linear constraints. In this paper, we consider the post-selection inference problem for linear models with linear equality and inequality constraints. We present methods for constructing the confidence intervals for the selected coefficients, which are chosen based on a Lasso-type estimator with linear constraints. These confidence intervals are interpreted as having desired coverage probability when
conditioned on the selected model. Simulations are conducted for different settings and the results show the validity of our method in providing valid confidence intervals after variable selection. We also applied our method to the UK Twins microbiome dataset to study the association between taxa compositional and Age.

**Bayesian Inference of Multivariate Mixed-effects Joint Models for Longitudinal-survival Data**

Lan Xu and Yangxin Huang
University of South Florida

Joint models of longitudinal and time-to-event data have received a lot of attention in epidemiological and clinical research under a linear mixed-effects model with normal assumption for a single longitudinal outcome and Cox proportional hazards model. However, those models-based analysis may not provide robust inference when longitudinal measurements exhibit skewness and/or heavy tails. In addition, the data collected are often featured by multivariate longitudinal measurements with an attempt to cope with correlation may lead to biased estimation. In this article, under longitudinal outcome and Cox proportional hazards model. Our proposed models and method are evaluated by simulation studies and an application to construct gray matter network modality and account for the heterogeneity in networks due to differences between subjects and networks of external modality.

**Session 68: Statistical Challenges and Advances in Biomarker Studies**

**Statistical Methods for Genetic Risk Prediction**

Yiming Hu1, Qiongshi Lu2, Mo Li3, Yixuan Ye1, Wei Jiang1 and Hongyu Zhao1

1Yale University
2University of Wisconsin Madison
3Yale University

Accurate prediction of disease risk based on genetic factors is an important goal in human genetics research and precision medicine. Advanced prediction models will lead to more effective disease prevention and treatment strategies. Despite the identification of thousands of disease-associated genetic variants through genome-wide association studies in the past decade, accuracy of genetic risk prediction remains moderate for most diseases, which is largely due to the challenges in both identifying all the functionally relevant variants and accurately estimating their effect sizes. In this presentation, we will discuss our methods that can better integrate other types of data, e.g. genome annotation and joint trait analysis, in genetic risk prediction for complex diseases. The usefulness of our methods will be demonstrated through their applications to a number of traits/diseases.

**New Directions in image-based analysis of biomarkers**

Debashis Ghosh, Xuhong Zhang, Fuyong Xing and Olivier Simon
Colorado School of Public Health

debashis.ghosh@ucdenver.edu

With the expansion in the use of microscopy technologies, there is now a plethora of data available in biomedical settings on cellular measurements at very high-resolution. This leads to considerations of ‘big data’ problems in which gigabytes of data are routinely being generated. We will describe ongoing efforts in machine learning-based analysis of cellular-level imaging datasets at the University of Colorado. A variety of motivating biomedical examples will be considered.

**Using Information Theory to Evaluate Surrogacy of Biomarkers**

Maria Carmen1, Hua Jin2, Qian Zhao3 and Ying Lu4

1Complutense University of Madrid
2South China Normal University
3Guangzhou Medical University
4Stanford University

The evaluation of new treatments is based on the choice of clinically meaningful clinical endpoints. However, observing such endpoints may require very long follow-up times or may be too expensive to measure. These might be avoided through replacing the true endpoints (T) by surrogate endpoints (S) or biomarkers that can be measured earlier or more cheaply than the true clinical endpoint of interest. Alonso and Molenberghs (2007,2008) redefined surrogacy in terms of the information content that S provides with respect to T. This information can be measured using information theory measures. They based their proposal on the Shannon entropy. However, there are cases where Shannon’s entropy function does not exist or not normally distributed. So we propose to extend their measure to a family of measures based on Havrda and Charvat entropy without that drawback. We define a model when T is a normally distributed, t-distributed or binary distributed clinical endpoints and S are longitudinal continuous biomarkers variables. We estimate the model unknown parameters and plugged into the proposed measures. Finally, we revisited an ophthalmologic case study to illustrate our methodology.

**Identifying Biomarker Networks From Multiple Modality Data**

Yuanjia Wang1, Shanghong Xie1 and Donglin Zeng2

1Columbia University
2University of North Carolina at Chapel Hill
yw2016@cucm.columbia.edu

The biomarker networks measured by different modalities of data (e.g., structural magnetic resonance imaging (sMRI), diffusion tensor imaging (DTI)) may share the same true unknown underlying biological model. In this work, we propose a node-wise biomarker graphical model to leverage the shared mechanism between multi-modality data to provide a more reliable estimation of the target network modality and account for the heterogeneity in networks due to differences between subjects and networks of external modality. Latent variables are introduced to represent the shared unobserved biological network and the information from the external modality is incorporated to model the distribution of the underlying network. The performance of the proposed method is demonstrated by extensive simulation studies and an application to construct gray matter brain atrophy network of Huntington’s disease by using sMRI data and DTI data. The estimated network measures are shown to be meaningful for predicting follow-up clinical outcomes in terms of patient stratification and prediction.

**Session 69: Modeling and inference for complex dependence in large datasets**

**How to Measure the Similarity of Data?**

Gabor Szekely
National Science Foundation
there is a constant $r > 0$ such that $d(S(x), S(x')) = r \times d(x, x')$ for all pairs of elements $x, x'$ in $M$. But how can we define the degree of similarity? Applications of this notion include fraud detection, fingerprint analysis, image analysis, DNA and RNA analysis, data mining, etc. This topic is closely related to measuring dependence: maximal similarity is the opposite of independence thus the degree of similarity is closely related to the degree of dependence. In Hilbert spaces where the metric is conditionally negative definite, distance correlation introduced by the speaker about ten years ago is a very good measure of dependence and also of similarity. But how can we measure similarity / dependence in more general metric spaces? Our journey starts with Thales’ similarity of triangles and ends with the earth mover’s correlation.

Multi-scale testing of independence and conditional independence for massive data
Shai Gorsky, Jiliang Mao and ♦ Li Ma
Duke University
li.ma@duke.edu

We generalize a recently introduced multi-scale divide-and-conquer framework for nonparametrically testing the independence between two random variables to the problems of (i) testing independence between random vectors possibly of large dimensions and (ii) testing conditional independence between two random variables conditional another. The proposed framework leads to testing procedures that achieves a computational complexity that scales approximately linearly with sample size, while achieving theoretical guarantees in terms of Type I error control without resorting to resampling. These features make our approach suitable for massive data sets with millions or more observations. Under this framework, the testing of independence and conditional independence is transformed into a multiple testing problem involving a collection of tests corresponding to dependency structures on different parts of the sample space at a cascade of resolution levels. The number of tests scale approximately linearly with sample size while the validity of the procedure can be addressed through multiple testing strategies. We compare both the statistical and computational efficiency with other state-of-the-art through extensive numerical studies.

Detect with BERET
♦ Duyeol Lee, Kai Zhang and Michael Kosorok
University of North Carolina at Chapel Hill
leeduyeol@gmail.com

Recently, the binary expansion testing framework was introduced to test the independence of two continuous random variables by utilizing symmetry statistics which are complete sufficient statistics for dependence. In this paper, we develop a new test through an ensemble method utilizing both the sum of squared symmetry statistics and distance correlation. Simulation studies suggest that the proposed method improves the power while preserving the clear interpretation of the binary expansion testing. We further extend this method to tests of independence of random vectors in arbitrary dimension. The proposed binary expansion randomized ensemble test (BERET) transforms the multivariate independence testing problem into a univariate one through random projections. The power of the proposed method is illustrated with many simulated and real data examples.

Essential Regression
♦ Xin Bing and Florentina Bunea
Cornell University
xb43@cornell.edu

We introduce the Essential Regression model, which provides an alternative to the ubiquitous K-sparse high dimensional linear regression on $p$ variables. While K-sparse regression assumes that only K components of the observable X directly influence Y , Essential Regression allows for all components of X to influence Y , but mediated through a K-dimensional random vector Z. The mediator Z is unobservable, and made interpretable via modeling assumption through which each component of Z is given the physical meaning of a small group of the X-variables.Formally, E-Regression is a new type of latent factor regression model, in which the unobservable factors Z influence linearly both the response Y and the covariates X. Its novelty consists in the conditions that give Z interpretable meaning as well as render the regression coefficients $\beta \in RK$ relating Y to Z - along with other important parameters of the model - identifiable. The interpretability of the latent factors Z in Essential Regression allows us to provide conceptually new inferential tools for regression in high dimensions. In particular, K-dimensional inference at the Z level is a viable alternative to existing approaches that benefits from the greater simplicity and lower dimensionality of the “essence” Z when compared to X. It is furthermore well known that inference performed in K-sparse regression - after consistent support recovery and estimation of K - is valid only for the large regression coefficients of Y on X, which makes this approach problematic in practice. In contrast, inference performed in regression at the lower resolution level given by Z is uniform over the space of the regression coefficients $\beta \in RK$, although it is performed after estimating consistently K and the subset of the X-variables that explain Z. For this we construct computationally efficient estimators of $\beta$, derive their minimax rate, and show that they are minimax-rate optimal in Euclidean norm for every sample size $n$. We show that the component-wise estimates of $\beta$ are asymptotically normal, with small asymptotic variance. This is a new addition to the literature in factor models, in which the inference problem is under-explored.Prediction of Y from X under E-Regression complements, in the low signal to noise ratio regime, the battery of methods developed for prediction under other factor model specifications. Similarly to other methods, it is particularly powerful when p is large, with further refinements made possible by the Essential Regression model specifications. E-Regression also provides a statistical framework for analysis in regression with clustered predictors, with or without overlap. This allows us to address possible inferential questions in post clustering-inference, and subsequently provide guidelines regarding the use and misuse of cluster averages as very popular dimension reduction devices in high dimensional regression.

Session 70: Contemporary Viewpoints on Improving Adaptive Phase I Clinical Trial Designs

How Many Patients Should We Enroll in an Adaptive Phase I Trial?
Thomas Braun
University of Michigan
tcombraun@umich.edu

In contrast with typical Phase III clinical trials, there is little existing methodology for determining the appropriate numbers of patients to enroll in adaptive Phase I trials. And, as stated by Dennis Lindley...
in a more general context, the simple practical question of "What size of sample should I take" is often posed to a statistician, and it is a question that is embarrassingly difficult to answer." Historically, simulation has been the primary option for determining sample sizes for Bayesian adaptive Phase I trials, and although useful, can be problematic and time-consuming when a sample size is needed relatively quickly. I will present two recent and computationally easy approaches for determining sample size based upon approximating posterior DLT rate distributions with Beta distributions. The operating characteristics of each approach are examined via simulation of realistic Phase I trial settings.

Phase I designs that allow for uncertainty in the attribution of adverse events

Alexia Iasonos and John Oquigley
1Memorial Sloan Kettering Cancer Center
2University Paris 6
lasonosa@mskcc.org

In determining dose limiting toxicities in Phase I studies, it is necessary to attribute adverse events (AE) to being drug related or not. Such determination is subjective and may introduce bias. We developed methods for removing or at least diminishing the impact of this bias on the estimation of the maximum tolerated dose (MTD). The proposed approach takes into account the subjectivity in the attribution of AE by using model-based dose escalation designs. It allows the investigators to strictly adhere to the protocol by recording all dose limiting toxicities and, yet, still allow expert opinion to prevent AE that are most likely not drug related from seriously compromising the exercise of identifying the correct MTD. The results show that gains can be achieved in terms of accuracy by recovering information lost to biases. These biases are a result of ignoring the errors in toxicity attribution. Theoretical results and simulation studies under small sample sizes will be presented.

Shift models for dose finding in partially ordered groups

Bethany Horton, Nolan Wages and Mark Conaway
1University of Virginia
nwages@virginia.edu

Limited options are available for dose finding clinical trials requiring group specific dose selection. While conducting parallel trials in each group is an accessible approach to the problem, this approach permits maximum tolerated dose (MTD) selection that does not align with clinically meaningful group order information. This talk describes the development of the two-stage continual reassessment method (CRM) for dose-finding studies involving three or more groups where group frailty order is known between some but not all groups, creating a partial order. This is an extension of the existing CRM shift model for two ordered groups. This method allows for dose selection by group, where MTD selection follows the known frailty order among groups. For example, if a group is known to be the most frail, the recommended MTD for this group should not exceed the MTD recommended for any other group. With limited alternatives for dose finding in partially ordered groups, this method is compared to two alternatives: 1) an existing method for dose finding in partially ordered groups proposed by Conaway (2017) and 2) independent trials for each group using the two-stage CRM. Simulation studies show that when ignoring information on group frailty, using independent CRM trials by group, 30% of simulations would result in MTD selection that is out of order between groups. In addition, the two-stage CRM for partially ordered groups selects the MTD more often and assigns more patients to the MTD compared to using independent CRM trials within each group. Simulation results for the proposed method and the Conaway method are similar. The proposed CRM for partially ordered groups ensures appropriate MTD order and improves accuracy of MTD selection over independent trials, while allowing for trial implementation that is computationally accessible.

Combining Data from a Phase I-II dose Escalation Study and a Follow-up Cohort

Joseph Koopmeiners, Kasha Mohammadi and Jeffrey Boatman
1University of Minnesota
2University of Texas
koopm007@umn.edu

The objective of phase I-II oncology trials is to identify the optimal dose for further study by considering the trade-off between efficacy and toxicity during dose-finding. In addition to identifying the optimal dose, it is also important to achieve an accurate estimate of the probability of efficacy and toxicity. Dose-finding trials are often followed by single-arm trials to further evaluate the optimal dose. Combining the data across these two sources has the potential to increase efficiency but could also introduce bias in the presence of inter-trial heterogeneity. In this talk, we explore various approaches to combining data between a phase I-II dose escalation trial and a follow-up cohort, including: naive approaches that use only the expansion cohort data or that pool the escalation and expansion cohort data; an estimation of the Uniformly Minimum Variance Unbiased Estimator (UMVUE) for estimating a binomial proportion from a two-stage design, and two Bayesian approaches based on commensurate priors and multisource exchangeability models. Simulation results suggest simply pooling the data performs more efficiently in the absence of inter-trial heterogeneity. However, the Bayesian method performed better in the presence of inter-trial heterogeneity. We illustrate the proposed estimators using data from a phase I-II clinical trial of a novel treatment for canine-hemangiosarcoma and a follow-up single arm trial.

Session 71: Advances in High-Dimensional Inference and Regression

An Empirical Bayes Approach for Selection Bias in Functional Data

Joshua Derenski, Yingying Fan and Gareth James
University of Southern California
derenski_joshua@gmail.com

Over the past few decades, functional data sets have become more common in many scientific fields (astronomy, ecology, etc.). As with other types of data, functional data often suffer from selection bias: the phenomenon that extreme estimates systematically overestimate (or underestimate) their mean values. For functional data, selection bias extends to functionals of curves. To that end, we present FEmBa: An empirical Bayes approach for handling selection bias in functional data. Using FEmBa, we obtain more accurate estimates of the mean functions of a series of curves (where each curve has an associated mean function), with greater improvement for “extreme” curves and in many cases functionals of those curves. We demonstrate this improvement through a series of simulations, and then apply FEmBa to the problem of estimating exoplanet sizes, and estimating the effect of human activity on water levels in several aquifers.

RANK: Large-Scale Inference with Graphical Nonlinear Knockoffs

Yingying Fan, Emre Demirkaya, Guorong Li and Jinchi Lv
Power and reproducibility are key to enabling refined scientific discoveries in contemporary big data applications with general high-dimensional nonlinear models. In this paper, we provide theoretical foundations on the power and robustness for the model-X knockoffs procedure introduced recently in Candès, Fan, Janson and Lv (2018) in high-dimensional setting when the covariate distribution is characterized by Gaussian graphical model. We establish that under mild regularity conditions, the power of the oracle knockoffs procedure with known covariate distribution in high-dimensional linear models is asymptotically one as sample size goes to infinity. When moving away from the ideal case, we suggest the modified model-X knockoffs method called graphical nonlinear knockoffs (RANK) to accommodate the unknown covariate distribution. We provide theoretical justifications on the robustness of our modified procedure by showing that the false discovery rate (FDR) is asymptotically controlled at the target level and the power is asymptotically one with the estimated covariate distribution. To the best of our knowledge, this is the first formal theoretical result on the power for the knockoffs procedure. Simulation results demonstrate that compared to existing approaches, our method performs competitively in both FDR control and power. A real data set is analyzed to further assess the performance of the suggested knockoffs procedure.

**Effective Joint Modeling through the Multivariate Debiased Lasso**

*Jacob Rhyne*

North Carolina State University [jdrhyne2@ncsu.edu](mailto:jdrhyne2@ncsu.edu)

In expression quantitative trait loci (eQTL) mapping, associations are identified between genetic variants, usually Single Nucleotide Polymorphisms (SNPs), and the expression of genes. One method of eQTL mapping is joint modeling, in which associations are identified between all SNPs and the expression of genes simultaneously through fitting one model. One well-known method of joint modeling is LORS, of Yang et al. (2013), which jointly models the expression of genes and SNPs while also accounting for hidden, unobserved factors that influence gene expression. While LORS has many positive attributes, the distribution of the estimate of the coefficients is unknown, thus statistical inference is limited. In this paper, we propose a new joint modeling method, MDLasso, that extends the debiased lasso of van de Geer et al. (2014) to a multivariate response. The proposed method adjusts for hidden factors, as in LORS, and produces an estimate for which inference can be performed. In simulation studies and analysis of HapMap eQTL data, it is shown that MDLasso with an appropriate multiple testing procedure successfully controls false discoveries while identifying a majority of true signals.

**Session 72: Subgroup identification techniques and practical applications in drug development**

*Ilya Lipkovich* and *Alex Dmitrienko*

Eli Lilly and Company [ilya.lipkovich@gmail.com](mailto:ilya.lipkovich@gmail.com)

This talk will provide a review of statistical methods dealing with exploratory subgroup analysis in clinical trials, as one of the key components of personalized medicine. This includes methods that can be applied in both early and late-phase clinical trials, as well as to observational studies (for some of the methods). A broad taxonomy of existing approaches to subgroup/biomarker identification will be given illustrating the key elements of principled data-driven subgroup evaluation using a case study.

**Subgroup Identification Through Confident Inference on Treatment Efficacy for Time-to-Event Outcomes**

*Yue Wei and Ying Ding*

University of Pittsburgh [yuw95@pitt.edu](mailto:yuw95@pitt.edu)

The uptake of targeted therapies has significantly changed the field of medicine. Instead of the “one-fits-all” approach, one aspect is to develop treatments that target a subgroup of patients. In this process, usually many markers need to be screened, and within each marker, it is necessary to infer treatment efficacy in subgroups and their combinations. For example, for a SNP that separates patients into three subgroups (AA, Aa and aa), one has to decide whether to target a single subgroup (e.g., aa) or a combination of subgroups (e.g., Aa, aa). In this research, we develop a simultaneous inference procedure to identify subgroups with enhanced treatment efficacy in clinical trials with time-to-event outcomes. Specifically, after establishing a suitable efficacy measure, we provide simultaneous confidence intervals, which appropriately adjust both within- and between-marker multiplicities, for comparing subgroups or combinations of subgroups. Realistic simulations are conducted using true genotype data and various efficacy scenarios to evaluate method performance. The method is then applied to the Age-Related Eye Disease Study (AREDS) data for subgroup identification. This approach provides a step toward confidently identifying patient subgroups in targeted therapy development.

**Subgroup Identification and Multi-population Tailoring Studies: Moving from Phase II to Phase III**

*Ying Zhang* and *Lei Shen*

Eli Lilly and Company [zhang.ying5@lilly.com](mailto:zhang.ying5@lilly.com)

When a candidate subgroup with differential treatment effect is identified in Phase II, multi-population tailoring studies can be conducted in Phase III. To evaluate treatment efficacy in the overall patient population as well as the subgroup while controlling the familywise type I error rate, an alpha-splitting approach can be employed in which alpha is allocated for sub- and overall populations. In this presentation, we investigate such alpha allocation in realistic scenarios in drug development using simulations under various overall treatment effect sizes, subgroup effect sizes, and sizes of subpopulation. We conclude that larger proportion of alpha should be allocated to overall testing when less differential effect is found in the subgroup. We also find that optimizing the alpha-allocation is especially important for the overall power (at least one success in the two populations) when the size of the subgroup is small or the treatment effect size is moderate in the overall population.

**Session 73: Recent Advances in joint modeling and dynamic prediction on longitudinal and survival data**

**Assessing predictive accuracy of survival regressions subject to non-independent censoring**

*Ming Wang* , *Qi Long* , *Chixiang Chen* and *Lijun Zhang*

Pennsylvania State University [ming.wang@pennstate.edu](mailto:ming.wang@pennstate.edu)

This talk will provide a review of statistical methods dealing with exploratory subgroup analysis in clinical trials, as one of the key components of personalized medicine. This includes methods that can be applied in both early and late-phase clinical trials, as well as to observational studies (for some of the methods). A broad taxonomy of existing approaches to subgroup/biomarker identification will be given illustrating the key elements of principled data-driven subgroup evaluation using a case study.
Bin Shi

Landmark analysis for the effects of longitudinal cholesterol longitudinal history. Finally, I will present a class of statistical models for competing risks, and predictive feature extraction from longitudinal data under CAR or NCAR for extensions. The performance of the proposed method is evaluated through extensive simulation studies and analysis of real data from the Critical Assessment of Microarray Data Analysis.

A class of models for dynamic prediction of survival using multivariate longitudinal data

Liang Li
University of Texas MD Anderson Cancer Center

I will first present a brief overview of dynamic prediction problems, including the real data application, model formulation, estimation, and validation. Then, I will discuss the analytical challenges often encountered in chronic kidney diseases, including multivariate longitudinal predictors of multiple types, nonlinear longitudinal trajectories, competing risks, and predictive feature extraction from longitudinal history. Finally, I will present a class of statistical models recently developed by our group that address these challenges.

Landmark analysis for the effects of longitudinal cholesterol profiles on the risk of CHDs

Bin Shi1, Peng Wei2 and Xuelin Huang2
1University of Texas
2University of Texas MD Anderson Cancer Center

Cardiovascular disease is the number one cause of mortality in both women and men in the United States. It is widely accepted that treatments that reduce cholesterol levels are beneficial in adults of approximate ages of 40 to 69 years, but any benefit in adults 70 years and older is still under debate. In this research, we applied landmark analysis combined with functional principal component analysis models to estimate coronary heart disease (CHD) risk profiles for longitudinal total cholesterol levels using data from the Framingham Heart Study. We found different risk patterns that may explain the overall trend: higher total cholesterol was associated with higher CHD risk in younger age range. However, in adults 70 years and older, higher cholesterol levels were not associated with higher CHD risk. Our analysis provides more crucial evidence to inform the ongoing debate, and demonstrates the importance of analyzing age-dependent effects of TC on CHD risks in a patient’s early age.

Session 74: Recent advances in modern nonparametric statistics with applications in ROC curve analysis

Nonparametric Mass Imputation for data integration
Sixia Chen

Nonprobability samples have been used frequently in practice due to the lack of sampling frame information, time or budget. Moreover by only using nonprobability samples without further adjustments may lead to biased results. Parametric mass imputation approaches have been developed in previous research, but the performances depend on the underlying parametric model assumptions. To overcome this issue, we propose nonparametric mass imputation for data integration. For low dimensional covariate, kernel smoothing approach is proposed. For relatively high dimensional covariate, generalized additive model is used for imputation. Asymptotic theories have been developed. Simulation studies as well as real application show the benefits of our proposed methods compared with parametric methods.

Assessing Partial Association between Ordinal Variables: A General Framework

1Dungang Liu1, Shaobo Li2 and Yan Yu1
1University of Cincinnati
2University of Kansas

deep learning, which can give us deeper insights into the association of variables and it only reflects linear association. In this paper, we propose a new framework for studying ordinal-ordinal partial association, by using surrogate residuals (Liu and Zhang, JASA, 2018) derived from fitting ordinal regression models. We justify that Y1 and Y2 are partially independent if and only if the corresponding surrogate residual variables are independent. Based on this theoretical result, we develop a general measure to quantify the size and shape of ordinal-ordinal partial association. As opposed to the polychoric correlation, our measure has the following properties: (1) its size reflects the strength of association for ordinal data, rather than the latent “data”; (2) it does not rely on the normality assumption or parametric methods.

Partial association measures the relationship between two variables Y1 and Y2 after adjusting a set of covariates X. It has remained unknown how to fully characterize such an association if both Y1 and Y2 are recorded on ordinal (or binary) scales. A classical approach is to use as a substitute the partial correlation between two latent continuous variables. This so-called polychoric correlation is inadequate, as it requires the bivariate normality of the latent variables and it only reflects linear association. In this paper, we propose a new framework for studying ordinal-ordinal partial association, by using surrogate residuals (Liu and Zhang, JASA, 2018) derived from fitting ordinal regression models. We justify that Y1 and Y2 are partially independent if and only if the corresponding surrogate residual variables are independent. Based on this theoretical result, we develop a general measure to quantify the size and shape of ordinal-ordinal partial association. As opposed to the polychoric correlation, our measure has the following properties: (1) its size reflects the strength of association for ordinal data, rather than the latent “data”; (2) it does not rely on the normality assumption or parametric models with the probit link, but instead it broadly applies to models with any link functions; and (3) it can capture non-linear associations and has potential to detect dependence of any complex structures. Our measure can be complemented by visualization methods, such as partial regression plots and P-P plots, which were otherwise unavailable for ordinal data. We stress that the focus of this paper is not on hypothesis testing, but quantification and visualization. Simulated and real examples demonstrate that our numerical and graphical assessment can reveal microstructure of partial association, which can give us deeper insights into the association of interest.

Bandwidth-free ROC Curve Estimation via Bernstein Polynomial

1Dongliang Wang1 and Xueya Cai2
1SUNY Upstate Medical University
2University of Rochester

The receiver operating characteristic (ROC) curve has been widely used for assessing the accuracy of a diagnostic test. Given its central impact over many summary measures, such as the area under the curve (AUC) and the Youden index, reliable ROC curve estimation remains highly desirable, despite of a variety of methods.
developed from parametric, semiparametric and nonparametric perspectives. Via smoothing the empirical ROC curve with Bernstein type polynomials, a novel ROC curve estimator is proposed, which is simple to calculate with no burden for bandwidth selection, and transformation invariant. Empirical results from a comprehensive simulation study are presented in order to evaluate the performance of the proposed estimator in comparison with many other existing ones in term of mean squared error (MSE) and to provide practical insights for selecting appropriate estimating methods for real data analysis.

Session 75: Recent Advances in Neuroimaging Analysis

Defining the resolution of optogenetic circuit mapping
♦ Shizhe Chen1, Ben Shababo, Karl Kilborn, Xinyi Deng, Johannes Friedrich, Hillel Adesnik and Liam Paninski
1University of California, Davis
szdcchen@ucdavis.edu
Circuit-mapping experiments combining whole-cell electrophysiology with two-photon optical stimulation of potentially presynaptic neurons have produced rich data on monosynaptic connectivity of neural circuits. However, mapping densely-packed presynaptic populations at cellular resolution has proven challenging, making the precise localization of connected neurons difficult. To interpret data resulting from these experiments, it is therefore critical to characterize the spatial resolution of stimulation. We develop a generative 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 a voltage-clamp observation. We develop efficient online Bayesian variational inference methods for tracking the posterior of the model parameters given the observed data. We use simulated and real data to characterize the resulting resolution limits, and compare the accuracy of the proposed inference methods against simpler baseline approaches.

Sparse Generalized Eigenvalue Problem with an Application to Neuroscience
Kean Ming Tan
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. Most of the existing methods and theory in the context of specific statistical models that are special cases of sparse GEP require restrictive structural assumptions on the input matrices. This talk will focus on a two-stage computational framework for solving the non-convex optimization problem resulting from 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 non-convex optimization perspective and propose the truncated Rayleigh flow method (Rifle) to estimate the leading generalized eigenvector, and show that it converges to a solution with the optimal statistical rate of convergence. Theoretically, our method significantly improves upon the existing literature by eliminating the structural assumptions on the input matrices. Numerical studies in the context of several statistical models are provided to validate the theoretical results. We then apply the proposed method to an electrocorticography data to understand how human brains recall and mentally rehearse word sequences.

Covariate Assisted Principal Regression for Covariance Matrix Outcomes
♦ Yi Zhao1, Bingkai Wang2, Stewart Mostofsky1, Brian Caffo1 and Xi Luo2
1Johns Hopkins University
2University of Texas
zhaoyi1026@gmail.com
Modeling variances in data has been an important topic in many fields, including in financial and neuroimaging analysis. We consider the problem of regressing covariance matrices on a vector of covariates, collected from each observational unit. The main aim is to uncover the variation in the covariance matrices across units that are explained by the covariates. This study introduces Covariate Assisted Principal (CAP) regression, an optimization-based method for identifying the components predicted by (generalized) linear models of the covariates. We develop computationally efficient algorithms to jointly search the projection directions and regression coefficients, and establish the asymptotic properties. Using extensive simulation studies, our method shows higher accuracy and robustness in coefficient estimation than competing methods. Applied to a resting-state functional magnetic resonance imaging study, our approach identifies meaningful findings and equips network-level interpretations.

Individualized Multilayer Tensor Learning with An Application in Imaging Analysis
♦ Xiwei Tang1, Xuan Bi2 and Annie Qu3
1University of Virginia
2University of Minnesota
3University of Illinois at Urbana-Champaign
xt4yj@virginia.edu
This work is motivated by multimodality breast cancer imaging data, which is quite challenging in that the signals of discrete tumor-associated microvesicles (TMVs) are randomly distributed with heterogeneous patterns. This imposes a significant challenge for conventional imaging regression and dimension reduction models assuming a homogeneous feature structure. We develop an innovative multilayer tensor learning method to incorporate heterogeneity to a higher-order tensor decomposition and predict disease status effectively through utilizing subject-wise imaging features and multimodality information. Specifically, we construct a multilayer decomposition which leverages an individualized imaging layer in addition to a modality-specific tensor structure. One major advantage of our approach is that we are able to efficiently capture the heterogeneous spatial features of signals that are not characterized by a population structure as well as integrating multimodality information simultaneously. To achieve scalable computing, we develop a new bi-level block improvement algorithm. In theory, we investigate both the algorithm convergence property, tensor signal recovery error bound and asymptotic consistency for prediction model estimation. We also apply the proposed method for simulated and human breast cancer imaging data. Numerical results demonstrate that the proposed method outperforms other existing competing methods.
Session 76: The Jiann-Ping Hsu Invited on Biostatistical and Regulatory Sciences

A measure and cut-point selection criterion for k-stage diseases using Kullback-Leibler divergence.  
\*Chen Mo, Hani Samawi, Jingjing Yin and Jing Kersey  
Georgia Southern University  
\texttt{cm06957@georgiasouthern.edu}

Recently, Kullback-Leibler (KL) divergence, which captures the disparity between two distributions, has been considered as a measure for determining the diagnostic performance of biomarkers. This study proposes a new measure generalized from KL divergence, which is named total Kullback-Leibler divergence using adjacent cells (TKLA), for diseases with multi-stage (k ≥ 2). The study investigates the performance of TKLA in medical diagnostics, including comprehensive measures of rule-in and rule-out potential. It also proposes an optimization criterion based on TKLA for cut-point selection for a multi-stage disease. Moreover, the geometric and probabilistic interpretations for TKLA are both made. Furthermore, the study illustrates the application of TKLA in real data.

Misclassification Simulation Extrapolation in survival analysis with AFT model  
\*Congjian Liu and Lili Yu  
Georgia Southern University  
\texttt{cl103124@georgiasouthern.edu}

The correct classification of variables is related to the accuracy of statistical model parameter estimation. The well know existing deviation correction method is Misclassification Simulation Extrapolation (MC-SIMEX) procedure. Survival analysis is the main part of time to event analysis and including Cox Proportional-Hazards model (COX model) and Accelerated Failure Time model (AFT model) two main aspects. We found that the COX models are well explored. However, bias-correction in AFT models with MC-SIMEX method still need be investigated, since some of the results of bias-correction are not well accepted. In this study, we use non-parametric extrapolation function instead of linear extrapolation function within AFT model to help us correct bias in AFT model with misclassification. In addition, we evaluated the performance of a modified version of the MC-SIMEX in non-parametric extrapolation function within AFT model.

Regularization in Accelerated Failure Time (AFT) models with frailty parameters  
\*Sarbesh Pandeya, Lili Yu, Hani Samawi and Xinyan Zhang  
Georgia Southern University  
\texttt{sp03459@georgiasouthern.edu}

Variable selection is one of the standard ways of conducting model selection in large scale data-sets. It is used in many studies especially in large multi-center clinical trials. One of the prominent methods in variable selection is the penalized likelihood which is both consistent and efficient. However, penalized selection in mixed effects models is significantly challenging because of the influence of random covariates. It is even more complicated when there is involvement of censoring as such issues may cause the equations for the maximum likelihood to not converge. Therefore, we proposed the penalized quasi-likelihood (PQL) approach to estimate the maximum likelihood and thereby introduced a sparsity-inducing adaptive penalty function that makes the selection on both fixed and frailty effects in censored survival data. We used the parametric accelerated failure time (AFT) models with frailty parameters and left censoring mechanism to develop the predictive model. We also compared our penalty function with other established procedures via their performance on accurately choosing the correct coefficients and shrinking the false estimates towards zero.

A Quantitative Assessment of Risk for Subgroup Pursuit in Clinical Trials  
\*Xinzhou Guo\textsuperscript{1}, Ruoshua Li\textsuperscript{2} and Xuming He\textsuperscript{1}  
\textsuperscript{1}University of Michigan  
\textsuperscript{2}University of Texas Health Science Center at Houston  
\texttt{xinzhoug@umich.edu}

In clinical studies, when to recommend or decide further pursuit of the most promising subgroup that we have observed from an existing trial is a very important question. It is well recognized that the observed treatment effect size of the best identified subgroup tends to be too optimistic and, therefore, any subgroup pursuit needs to be made with a careful statistical consideration. In this paper, we address the issue of bias in subgroup pursuit and provide a quantitative screening measure that can be used in the decision-making of subgroup pursuit. In addition, we propose a confidence bound on the best subgroup treatment effect from clinical data. The proposed quantitative analysis is model-free, transparent and easy to compute, and can help make a better-informed decision of subgroup pursuit in clinical trials.

Session 77: High Dimensional Statistical Modelling of Genetics/Genomics

A Kernel-Based Neural Network for High-dimensional Genetic Risk Prediction Analysis  
Xiaoxi Shen, Xiaoran Tong and \*Qing Lu  
Michigan State University  
\texttt{qlu@epl.msu.edu}

Artificial intelligence (AI) is a thriving research field with many successful applications in areas such as imaging and speech recognition. Neural-network-based methods, such as deep learning, play a central role in the modern AI technology. While neural-network-based methods also hold great promise for genetic research, the high-dimensionality of genetic data and complex relationships between genetic variants and disease outcomes create tremendous analytic challenges. To address these challenges, we propose a kernel-based neural network (KNN) method. KNN inherits the high-dimensional feature from classical linear mixed models and the non-linear and non-additive features from neural networks, and is designed for high-dimensional genetic risk prediction analysis. KNN summarizes a large number of genetic variants into kernel matrices and uses the kernel matrices as input matrices. Based on the kernel matrices, KNN builds a feedforward neural network to model the complex relationship between genetic variants and a disease outcome. Minimum norm quadratic unbiased estimation (MINQUE) is implemented in KNN to make parameter estimation feasible. Through theoretical proof and simulations, we demonstrate that KNN can attain lower average prediction error than LMM. Finally, we illustrate KNN by an application to the sequencing data from the Alzheimer’s Disease Neuroimaging Initiate.

Genetic Heritability and Coefficients of Determination  
Dabao Zhang  
Purdue University  
\texttt{zhangdb@purdue.edu}

Genetic heritability is attached to the coefficient of determination for underlying statistical models. However, its calculation from high-throughput genomic data is challenged by the high-dimensionality of genetic data.
issue, and its definition for non-Gaussian models is challenged by lack of proper coefficients of determination. Here we will investigate and provide our solution to the definition and calculation of genetic heritability based on recent advances in statistics.

High Dimensional Mediation Analysis with Applications to Causal Gene Identification
Qi Zhang
University of Nebraska
qi.zhang@unl.edu

Mediation analysis has been a popular framework for elucidating the mediating mechanism of the exposure effect on the outcome. Previous literature in causal mediation primarily focused on the classical settings with univariate exposure and univariate mediator, with recent growing interests in high dimensional mediator. In this paper, we study the mediation model with high dimensional exposure and high dimensional mediator, and introduce two procedures for mediator selection, MedFix and MedMix. MedFix is our new application of adaptive lasso with one additional tuning parameter. MedMix is a novel mediation model based on high dimensional linear mixed model, for which we also develop a new variable selection algorithm. Our study is motivated by the causal gene identification problem, where causal genes are defined as the genes that mediate the genetic effect. For this problem, the genetic variants are the high dimensional exposure, the gene expressions the high dimensional mediator, and the phenotype of interest the outcome. We evaluate the proposed methods in extensive real data driven simulations, and apply them to causal gene identification from a mouse f2 dataset for diabetes study. We show that the mixed model based approach leads to higher accuracy in mediator selection and mediation effect size estimation, and is more reproducible across independent measurements of the response and more robust against model misspecification.

Statistical modeling of genomics and epigenomics big-data reveals human disease mechanisms
Jianrong Wang
Michigan State University
wangj164@msu.edu

Advances in population genetics, such as genome-wide association studies (GWAS), have made substantial progress to catalog which individual genetic variants are associated with specific human diseases. But we still do not know how these genetic variants, especially non-coding variants, lead to observed phenotypes. Therefore, there is a significant need to elucidate the underlying mechanisms by which the effects of individual genetic variants propagate to organism phenotypes through intermediate molecular disruptions, such as dysregulation of gene expression, interruption on chromatin structure and perturbation of biological pathways. In recent years, high-throughput next-generation sequencing techniques have revolutionized functional genomics research by generating large panels of genome-wide ‘omic’ datasets (e.g. epigenomics and transcriptomics). These abundant high-throughput datasets allow comprehensive annotations of regulatory elements and networks in different cell-types/tissues, which play essential roles for the purpose of delineating the molecular mechanisms linking individual genetic variants to complex human disease. Due to the ‘big data’ challenges, efficient and robust statistical models and machine learning algorithms are needed to integrate heterogeneous noisy functional genomics datasets. I will discuss several models we developed to dissect gene regulation systems using hierarchical graphical models. Major topics include discovery of new regulatory elements, 3D chromatin interactions, and combinatorial regulatory grammar. I will demonstrate how these inferences can be leveraged to identify causal regulatory SNPs and pathways in human diseases. The developed computational algorithms and the systems-level predictions will provide improved insights on gene regulation, chromatin architecture and disease mechanisms, leading to novel genomics-based diagnostics and therapeutics.

Session 78: Contributed Session 2

A Bayesian adaptive marker-stratified design with customized hierarchical modeling
• Yong Zhang\textsuperscript{1}, Beibei Guo\textsuperscript{2}, Yan Han\textsuperscript{3}, Sha Cao\textsuperscript{4} and Chi Zhang\textsuperscript{4}
\textsuperscript{1}Indiana University
\textsuperscript{2}Louisiana state university
\textsuperscript{3}Indiana university
\textsuperscript{4}Indiana university

It is well known that the treatment effect of a molecularly targeted agent (MTA) may vary dramatically, depending on each patients’ biomarker profile. Therefore, for a clinical trial evaluating MTA, it is more reasonable to evaluate its treatment effect within different marker subgroups rather than evaluating the average treatment effect for the overall population. The marker-stratified design (MSD) provides a useful tool to evaluate the subgroup treatment effects of MTAs. Under the Bayesian framework, the beta-binomial model is conventionally used under the MSD to estimate the response rate and test the hypothesis. However, this conventional model ignores the fact that the biomarker used in the MSD is in general predictive only for the MTA. The response rates for the standard treatment can be approximately consistent across different subgroups stratified by the biomarker. In this paper, we proposed a Bayesian hierarchical model incorporating this biomarker information into the consideration. The proposed model uses a hierarchical prior to borrow strength across different subgroups of patients receiving the standard treatment and therefore improve the efficiency of the design. The prior informativeness is determined by solving a “customized” equation reflecting the physician’s professional opinion. We developed a Bayesian adaptive design based on the proposed hierarchical model to guide the treatment allocation and test the subgroup treatment effect as well as the predictive marker effect. Simulation studies demonstrate that the proposed design yields desirable operating characteristics and outperforms the existing designs.

Bayesian propensity score analysis for multi-treatment groups
• Meijing Wu andHongwei Wang
Abbvie, Inc.
meijing.wu@abbvie.com

From early in drug development to life-cycle management, real world evidence (RWE) generated from real world data (RWD) is playing an increasingly important role for internal informed decision-making and external value demonstration. Due to its non-interventional nature, confounding bias is a key challenge needed to be addressed when analyzing RWD. Propensity score has been proved to be an effective means of dealing with large numbers of confounding covariates in observational studies especially for the scenario of binary treatment. The method includes treatment model in the first step and outcome model through subclassification, regression, weighting and matching in the second step. However, propensity score did not take the uncertainty of estimated propensity score into consideration and could not incorporate prior information in treatment effect estimation. Introducing Bayesian approach in
propensity score analysis could address the above problems. There had been several studies applying Bayesian approach in propensity score analysis since 2008. However, most of the studies were only focused on binary treatment. Bayesian approach in multi-treatment propensity score analysis remains unclear. This study aims to explore alternative methods of applying Bayesian approach in multi-treatment propensity score analysis, including both nominal and ordinal categorical treatments, and find out the optimal method for controlling confounding bias in multiple treatment groups. We proposed an “intermediate Bayesian generalized propensity score analysis” with Bayesian treatment model combined with conventional outcome model via regression, subclassification or weighting methods. We also proposed a “two-step Bayesian generalized propensity score analysis” with Bayesian treatment model in the first step combined with Bayesian regression as the Bayesian outcome model in the second step. A simulation study was conducted to evaluate the performance of two methods and practical guidance on optimal method was also provided. The following disclosure will be included in the presentation: The support of this presentation was provided by AbbVie. AbbVie participated in the review and approval of the content. Meijing Wu, Hongwei Wang are employees of AbbVie, Inc. The study was conducted when Meijing Wu studied at Secondary Military Medical University under the supervision and guidance of Jia He and Cheng Wu.

**A two-part linear mixed model with shared random effects for longitudinal microbiome compositions**

Yongli Han and Danping Liu
National Cancer Institute yongli.han@nih.gov

Longitudinal microbiome studies have been widely used to unveil the dynamics in the complex host microbial ecosystems. Modeling the longitudinal microbiome compositional data, which is semi-continuous in nature, is challenging in several aspects: the over-abundance of zeros, the heavy skewness of non-zero values that are bounded in (0;1), and the dependence between the binary and non-zero parts. To deal with these challenges, we first extended the work of Chen and Li (2016) and proposed a two-part zero-inflated Beta regression model with shared random effects (ZIBR-SRE), which characterize the dependence between the binary and the continuous parts. Besides, the microbiome compositional data have unit-sum constraint, indicating the existence of negative correlations among taxa. As ZIBR-SRE models each taxon separately, it does not satisfy the sum-to-one constraint. We then proposed a two-part linear mixed model (TPLMM) with shared random effects to formulate the log-transformed standardized relative abundances rather than the original ones. Such transformation is called “additive logistic transformation”, initially developed for cross-sectional compositional data. We extended it to analyze the longitudinal microbiome compositions and showed that the unit-sum constraint can be automatically satisfied under the TPLMM framework. Model performances of TPLMM and ZIBR-SRE were compared with existing methods in simulation studies. Under settings adopted from real data, TPLMM had the best performance and is recommended for practical use. An oral microbiome application further showed that TPLMM and ZIBR-SRE estimated a strong correlation structure in the binary and the continuous parts, suggesting models without accounting for this dependence would lead to biased inferences.

**Bayesian Analysis of Historical Data for Equivalence Acceptance Criteria under Nested Mixed Models**

Qianqiu Li, Jyh-Ming Shoung and Bill Pikounis

Janssen Research and Development, LLC
QLi16@its.jnj.com

Borrowing historical data in design and analysis of comparability studies has been increasingly adopted in pharmaceutical industries. In two-group comparisons of means or regression line slopes for manufacturing, stability or analytical testing studies involving one random factor nested within group, one or more historical data sets were evaluated using Bayesian approaches to help maintain statistical test power with small samples and to integrate data-related uncertainties. The risk-based posterior coverage intervals were derived to justify equivalence acceptance criteria of mean or slope differences. Model assumptions and commensurate priors were incorporated for incompleteness and adaptive use of historical data.

**An integrative sparse boosting analysis of cancer genomic commonality and difference**

Yifan Sun1, Yu Jiang2, Zhengyang Sun1, Yang Li3 and Shuangge Ma3

1Renmin University of China
2University of Memphis
3Yale University

sunyifan1984@163.com

In cancer research, high-throughput profiling has been extensively conducted. In recent studies, the integrative analysis of data on multiple cancer patient groups/subgroups has been conducted. Such analysis has the potential to reveal the genomic commonality as well as difference across groups/subgroups. However, in the existing literature, methods with a special attention to the genomic commonality and difference are very limited. In this study, we develop a novel estimation and marker selection method based on the sparse boosting technique to address the commonality/difference problem. In terms of technical innovation, a new penalty and computation of increments are introduced. The proposed method can also effectively accommodate the grouping structure of covariates. Simulation shows that it can outperform direct competitors under a wide spectrum of settings. The analysis of two TCGA (The Cancer Genome Atlas) datasets is conducted, showing that the proposed analysis can identify markers with important biological implications and have satisfactory prediction and stability.

**Session 79: Recent advance in precision medicine methods**

**Comparing Multiple Adaptive Interventions under General SMART Designs**

Xiaobo Zhong1, Bin Cheng2, Min Qian2 and Ying Kuen Cheung2

1Icahn School of Medicine at Mount Sinai
2Columbia University

bc2159@columbia.edu

A first step in identifying the best adaptive intervention (AI) embedded in a sequential multiple assignment randomized trial (SMART) is to make global comparisons among the AIs. However, the estimators of the AI values under a SMART design are generally correlated, and the correlation varies by design structures, imposing challenges in making global comparisons among the AIs. In this paper, two methods are proposed. One is a Wald test for global difference among the AIs, which yields a gate-keeping method to identify the best AI and a theoretically justified sample size formula under SMART. The other is a construction of asymptotic simultaneous confidence intervals for comparing with the true best AI embedded in a given SMART design. The proposed methods are...
Constructing Stabilized Dynamic Treatment Regimes for Censored Data

Yingqi Zhao, Ruqing Zhu, Guanhua Chen and Yingye Zheng

1Fred Hutchinson Cancer Research Center
2University of Illinois at Urbana-Champaign
3University of Wisconsin Madison

Stabilized dynamic treatment regimes are sequential decision rules for individual patients that not only adapt throughout the disease progression but also remain consistent over time in format. The estimation of stabilized dynamic treatment regimes becomes more complicated when the clinical outcome of interest is a survival time subject to censoring. To address this challenge, we propose two novel methods: censored shared-Q-learning and censored shared-O-learning. Both methods incorporate clinical preferences into a qualitative rule, where the parameters indexing the decision rules are shared across different stages and estimated simultaneously. We use extensive simulation studies to demonstrate the superior performance of the proposed methods. The methods are further applied to the Framingham Study to derive treatment rules for cardiovascular disease.

Stochastic Tree Search for Estimating Optimal Dynamic Treatment Regimes

Lu Wang and Yilan Sun

University of Michigan

A dynamic treatment regime (DTR) is a sequence of decision rules that adapt to time-varying state of an individual. Interpretable DTRs are desired for human experts to understand and implement. We develop a stochastic tree-based reinforcement learning method, ST-RL, for estimating optimal DTRs in a multi-stage multi-treatment setting utilizing data from either randomized trials or observational studies. At each stage, ST-RL constructs a decision tree by first modeling counterfactual outcome using Bayesian nonparametric regression models and then stochastically search for the optimal tree using Markov Chain Monte Carlo algorithm. The method is implemented in a backward inductive fashion through multiple decision stages. Compared to existing methods, ST-RL delivers DTRs with better interpretability, and does not require explicit working model specification. Moreover, ST-RL has stable and outstanding performance with moderately high dimensional data. We illustrate the performance of ST-RL through simulation studies, and also an application on esophageal cancer data collected from 1170 patients at MD Anderson Cancer Center from 1998 to 2012.

On Estimation of Optimal Treatment Regimes for Maximizing Quality Adjusted Survival Time

Sun Hat, Ashkan Ertefaie and Brent Johnson

University of Rochester

A treatment regime is a decision rule that maps observed individual prognostic information to a treatment option. There is increasing interest in finding optimal treatment regimes, which determine a sequence of treatments to optimize an expected clinical outcome, e.g., survival time. In many chronic diseases and for terminally-ill patients, extending the survival time may not be the only goal of the therapeutic strategy because it fails to take patient’s quality of life into consideration. This motivates us to focus on the quality adjusted survival lifetime (QAL) as an outcome of interest. Specifically, our goal is to find optimal treatment regimes that maximize some specific functions of QAL survival probability and one typical example we consider is the restricted qualify adjusted lifetime (RQAL). Standard survival analysis methods lead to biased estimation because of induced informative censoring which is a result of rescaling the original survival times. In this paper, we overcome this issue by developing a weighting estimating function approach that enables us to estimate the optimal regimes in a restricted functional class and provides consistent estimation of the value function. Asymptotic properties of the proposed estimator are established under suitable regularity conditions, and simulation studies are conducted to evaluate the finite sample performance of the proposal. The method is illustrated by application to a Ceftriaxone trial with ALS patients.

Session 80: Structure learning and statistical inference for complex data analysis

Nonparametric conditional density estimation based on pooled biomarker assessments

Xichen Mou, Dewei Wang and Joshua Tebbs

University of South Carolina
deweiwang@stat.sc.edu

In biomarker studies, when resources (e.g., the budget and/or the number of specimens) are limited, pooling specimens to take measurements of the biomarker’s concentration level is often an alternative means. This article develops a kernel-based regression estimator of a biomarker level’s density when a continuous covariate is available for each specimen but the biomarker is measured in pools with measurement errors. Consistency and asymptotic normality of our estimator is established. The rates of convergence depend on the tail behavior of the characteristic functions of the measurement error and the biomarker level. Simulation studies demonstrate the practical advantages of our method when comparing to the existing work. We further illustrate our method via a Polyfluorochemical data set.

Generalized probabilistic principal component analysis of correlated data

Mengyang Gu and Weining Shen

1Johns Hopkins University
2University of California, Irvine

Principal component analysis (PCA) is a well-established tool in machine learning and data processing. The principal axes in PCA were shown to be equivalent to the maximum marginal likelihood estimator of the factor loading matrix in a latent factor model for the observed data, assuming that the latent factors are independently distributed as standard normal distributions. However, the independence assumption may be unrealistic for many scenarios such as modeling multiple time series, spatial processes, and functional data, where the outcomes are correlated. In this paper, we introduce the generalized probabilistic principal component analysis (GPPCA) to study the latent factor model for multiple correlated outcomes, where each factor is modeled by a Gaussian process. Our method generalizes the previous probabilistic formulation of PCA (PPCA) by providing the closed-form maximum marginal likelihood estimator of the factor loadings and other parameters. Based on the explicit expression of the precision matrix in the marginal likelihood that we derived, the number of the computational op-
The two-phase design is a cost-effective sampling strategy to evaluate the effects of covariates on an outcome when certain covariates are too expensive to be measured on all study subjects. Under such a design, the outcome and inexpensive covariates are measured on all subjects in the first phase and the first-phase information is used to select subjects for measurements of expensive covariates in the second phase. Previous research on two-phase studies has focused largely on the inference procedures rather than the design aspects. We investigate the design efficiency of the two-phase study, as measured by the semiparametric efficiency bound for estimating the regression coefficients of expensive covariates. We consider general two-phase studies, where the outcome variable can be continuous, discrete, or censored, and the second-phase sampling can depend on the first-phase data in any manner. We develop optimal or approximately optimal two-phase designs, which can be substantially more efficient than the existing designs. We demonstrate the improvements of the new designs over the existing ones through extensive simulation studies and two large medical studies.

Learning Optimal Individualized Decision Rules with Risk Control

With the emergence of precision medicine, estimation of optimal individualized decision rules (IDRs) has attracted tremendous attention in many scientific areas. Most existing literature has focused on finding optimal IDRs that can maximize the expected outcome. The proposed criteria take tail behaviors of the outcome into consideration, and thus the resulting optimal IDRs are more likely to reduce adverse events. The optimal IDRs under our criteria can be substantially more efficient than the existing designs. We demonstrate the improvements of the new designs over the existing ones through extensive simulation studies and two large medical studies.

Session 81: Recent Advances of Statistical Modeling in Biomedical Research

Optimal designs of two-phase studies

Ran Tao

R. Tao

The two-phase design is a cost-effective sampling strategy to evaluate the effects of covariates on an outcome when certain covariates are too expensive to be measured on all study subjects. Under such a design, the outcome and inexpensive covariates are measured on all subjects in the first phase and the first-phase information is used to select subjects for measurements of expensive covariates in the second phase. Previous research on two-phase studies has focused largely on the inference procedures rather than the design aspects. We investigate the design efficiency of the two-phase study, as measured by the semiparametric efficiency bound for estimating the regression coefficients of expensive covariates. We consider general two-phase studies, where the outcome variable can be continuous, discrete, or censored, and the second-phase sampling can depend on the first-phase data in any manner. We develop optimal or approximately optimal two-phase designs, which can be substantially more efficient than the existing designs. We demonstrate the improvements of the new designs over the existing ones through extensive simulation studies and two large medical studies.
constrained set. Simulation studies and a real data application are used to demonstrate the robust performance of our methods.

**Session 82: Recent Methods for Complex Survival Data**

**Quantile Association Regression on Bivariate Survival Data**  
Jing-Li Chen, Yan Cheng, Ying Ding and Ruosha Li

1National Institute of Environmental Health  
2University of Pittsburgh  
3University of Texas Health Science Center at Houston  

The association between two event times is of scientific importance in various fields. The local association measures capture the dynamic pattern of association and thereby provide much richer information than the global association measures. Due to the heterogeneity of the population, it is desirable to examine the degree to which local association depends on different characteristics of the population. In this work, we adopt a novel quantile-based local association measure, which is free of marginal distributions, and propose a conditional quantile association regression model to allow covariate effects in the local association analysis for bivariate survival data. An estimating equation for the quantile association coefficients is constructed on the basis of the relationship between this quantile association measure and the conditional copula. The asymptotic properties for the resulting estimators are rigorously derived. In quantile regression and association analyses, one difficulty is the covariance estimation for which the true value involves the unknown density functions. To avoid estimating density functions, we extend the induced smoothing idea to our proposed estimators in obtaining the covariance matrix. The proposed estimators and inference procedure are evaluated through simulations, and applied to an Age-related Macular Degeneration (AMD) dataset, where we explore the association between AMD progression times in the two eyes of the same patient.

**A Varying-Coefficient Generalized Odds Rate Model with Time-Varying Exposure**  
Jie Zhou, Jiajia Zhang, Alexander Mclain, Wenbin Lu and James Hardin

1University of South Carolina  
2University of South Carolina  
3North Carolina State University  

Varying-coefficient models have become a common tool to determine whether and how the association between an exposure and an outcome changes over a continuous measure. These models are complicated when the exposure itself is time-varying and subjected to measurement error. For example, it is well known that longitudinal physical fitness has an impact on cardiovascular disease (CVD) mortality. It is not known, however, how the effect of longitudinal physical fitness on CVD mortality varies with age. In this paper, we propose a varying-coefficient generalized odds rate model that allows flexible estimation of age-modified effects of longitudinal physical fitness on CVD mortality. In our model, the longitudinal physical fitness is measured with error and modeled using a mixed effects model, and its associated age-varying coefficient function is represented by cubic B-splines. An expectation-maximization (EM) algorithm is developed to estimate the parameters in the joint models of longitudinal physical fitness and CVD mortality. A modified pseudo-adaptive Gaussian Hermite quadrature method is adopted to compute the integrals with respect to random effects involved in the E-step. The performance of the proposed method is evaluated through extensive simulation studies and is further illustrated with an application to cohort data from the Aerobic Center Longitudinal Study.

**Gene-based Association Test for Bivariate Survival Data via Functional Regression with Copula Models**  
Ying Ding

Department of Biostatistics, University of Pittsburgh  

Several gene-based association tests for time-to-event traits have been proposed recently, to detect whether a gene region (containing multiple variants), as a set, is associated with the survival outcome. However, for bivariate survival outcomes, to the best of our knowledge, there is no statistical method that can be directly applied for gene-based association analysis. Motivated by a genetic study to discover gene regions associated with the progression of a bilateral eye disease, Age-related Macular Degeneration (AMD), we implement a novel functional regression method under the copula framework. Specifically, the effects of variants within a gene region are modeled through a functional linear model, which then contributes to the marginal survival functions within the copula. Generalized score test and likelihood ratio test statistics are derived to test for the association between bivariate survival traits and the genetic region. Extensive simulation studies are conducted to evaluate the type-I error control and power performance of the proposed approach, with comparisons to several existing methods for a single survival trait, as well as the marginal Cox functional regression model using the robust sandwich estimator for bivariate survival traits. Finally, we apply our method to a large AMD study, the Age-related Eye Disease Study (AREDS), to identify gene regions that are associated with AMD progression. The method has been implemented and added into a newly developed R package CopulaCenR.

**Joint analysis of interval-censored recurrent events and survival data**  
Guanglei Yu, Yang Li, Liang Zhu, Hai Zhao, Jianguo Sun and Leslie Robison

1Eli Lilly and Company  
2University of North Carolina at Charlotte  
3University of Texas Health Science Center at Houston  
4Zhongnan University of Economics and Law  
5University of Missouri  
6St. Jude Children’s Research Hospital  

Joint analysis of interval-censored recurrent events and survival data has been attracting increasing attention. For the relationship between those two types of data, however, many existing approaches either specify some dependence structure with fixed distributions or assume independence given some covariates. Such assumptions may not be realistic and can result in misleading conclusions. With respect to covariate effects, previous work predominantly considers either additive or multiplicative effects alone to the underlying mean or hazard functions. We present a joint analysis approach that is not restricted to a distribution or structured correlation assumptions. Both additive and multiplicative covariate effects are simultaneously considered with respect to the mean function of recurrent events. For the survival data, the accelerated failure time model is employed to directly characterize covariate effects on survival times. For estimation of regression parameters, estimating equation-based procedures are developed and asymptotic properties of the proposed estimates are established. A model-checking procedure is also provided. An extensive simulation study is conducted through.
to assess finite sample performance of the proposed procedures, and the method is applied to a motivating data set from the Childhood Cancer Survivor Study.

Session 83: New statistical methods for big data in biomedical research

Combinatorial Inference

Junwei Lu¹, Matey Neykov² and Han Liu³
¹Harvard University
²Carnegie Mellon University
³Northwestern University

junweilu@hsph.harvard.edu

We propose the combinatorial inference to explore the global topological structures of graphical models. In particular, we conduct hypothesis tests on many combinatorial graph properties including connectivity, hub detection, perfect matching, etc. Our methods can be applied to any graph property which is invariant under the deletion of edges. On the other side, we also develop a generic minimax lower bound which shows the optimality of the proposed method for a large family of graph properties. Our methods are applied to the neuroscience by discovering hub voxels contributing to visual memories.

CORALS: Co-clustering analysis via regularized alternating least squares

Gen Li
Columbia University
gli2521@cumc.columbia.edu

Multi-way tensor array data are frequently encountered in real applications. Oftentimes, it is of interest to find coherent co-clusters consisting of subsets of features in each mode within a tensor. Co-clusters identify feature groups in different modes simultaneously and carry important interpretation. In this talk, we will introduce a new method for identifying co-clusters based on a regularized alternating least squares approach. We further generalize the method to deal with multi-way tensors with non-Gaussian entries. We will demonstrate the superior performance of the method using simulations and a range of real applications.

CESME: Cluster Analysis with Latent Semiparametric Mixture Models

Wen Zhou¹, Lysou Zhang¹, Lulu Wang² and Hui Zou³
¹Colorado State University
²Gilead Sciences, Inc.
³University of Minnesota
rcizw@stat.colostate.edu

Model-based clustering is one of the most popular statistical approaches from unsupervised learning and has been widely employed in practice for exploratory analysis, data visualization, sub-community identification, and quality control. Regardless of its wide applicability, the traditional distributional assumption such as Gaussianity is too stringent to be validated in general, and therefore prevents the model-based clustering to be used for data with complex distributions, such as highly skewness. In this paper, we propose a flexible semiparametric latent model to cluster multivariate data deviated from Gaussian. The model assumes that the observed random vectors are obtained from unknown monotone transformations of latent variables governed by a Gaussian mixture distribution. The identifiability of the proposed model is carefully studied. An alternating maximization procedure is developed to estimate the proposed model, whose convergence property is investigated by using finite-sample analysis. An interesting transition phenomenon of the convergence for the proposed algorithm, which is due to the presence of the unknown transformations, is explored and provides guidance on the design of the algorithm. The proposed method is also numerically assessed through extensive simulations and has demonstrated superior performance compared to most of the contemporary competitors.

Robust high-dimensional asymmetric data analysis with non-convex mean regression

Bin Luo and Xiaoli Guo
University of North Carolina at Greensboro
b.luo@uncg.edu

Asymmetry along with heteroscedasticity or contamination often occurs with the growth of data dimensionality. In high-dimensional data analysis, such irregular settings are usually overlooked for the theoretical convenience. In this paper, we study high-dimensional mean regression using Penalized Robust Approximated quadratic M-estimators (PRAM), in general irregular settings such as random errors are lack of symmetry and homogeneity, or the regressors are lack of sub-Gaussian assumption. To reduce the bias caused by irregular random errors, the PRAM estimator uses a family of loss function with flexible robustness and diverging parameters to approximate the mean function from the traditional quadratic loss. We first show that, in the ultra-high dimension setting where the dimensionality can grow almost exponentially with the sample size, with certain mild conditions on the covariate and loss functions with diverging parameters, the PRAM estimators have local estimation consistency at the minimax rate enjoyed by LS-Lasso. We further establish that the PRAM estimators with an appropriate non-convex penalty in fact agree with the local oracle solution with correct support and thus obtain the oracle property in high-dimensional setting by exploring the result in the low-dimensional case. In simulation, we numerically demonstrate the performances among six PRAM estimators using three types of loss functions all equipped with diverging parameters (the Huber loss, Tukey’s biweight loss and Cauchy loss) and two types of penalty functions (the Lasso and MCP penalties). Our simulation studies and NCI-60 data analysis exhibit satisfactory finite sample performance of the PRAM estimators under general irregular settings.

Session 84: Case Studies and Methods for Learning and Improving Healthcare Through EHRs

Repurposing Statistical Tools to Improve Data Quality for Time-varying Measurements in EHRs

Qingxia Chen and Trent Rosenbloom
Vanderbilt University Medical Center
cindy.chen@vanderbilt.edu

Electronic Health Records (EHR) systems, fueled by recent federal provisions in the HITECH Act, have been increasingly implemented at US hospitals. Huge amounts of longitudinal and detailed patient information, including lab tests, medications, disease status, and treatment outcome, have been accumulated and are available electronically. These large clinical databases are valuable and cost-effective data sources for clinical and translational research. Dense and irregularly recorded vital signs and lab measurements are great components of EHRs with potential error inputs. Manual evaluation is time- and cost-consuming and hence infeasible if not impossible. We proposed a residual-based statistical approach to dynamically identify potential error-prone measurements so that clinical
practitioners or patients themselves can either confirm or correct the records. The key feature of the method is to learn and update the model from the interactive response from the practitioners or patients. We applied the proposed method to Growth Chart Study using EHR data.

PIE: A prior knowledge guided estimation method for bias reduction in EHR-based research

♦ Jing Huang1, Rui Duan1, Rebecca Hubbard1, Yonghui Wu2, Jason Moore1, Hua Xu3 and Yong Chen
1University of Pennsylvania  
2University of Florida  
3University of Texas at Houston Health Science Center

Abstract: Electronic health records (EHR) contain a large amount of health information, including diagnoses, laboratory tests, and prescription medications. They provide an efficient and wide reaching source for clinical and biomedical research. However, the complex and inconsistent nature of EHR data bring additional challenges for such research. In particular, automated phenotyping algorithms, which extract patients’ disease, treatment, and response information from EHR using both structured and unstructured data through advanced informatics technologies, may create misclassification or measurement error due to limited sensitivity and specificity of the algorithms. Such errors in EHR-derived phenotypes can lead to systematic bias, substantially inflate type I error, and diminish statistical power, which ultimately lead to low reproducibility of EHR-based research findings. In this study, we propose a novel prior knowledge guided integrated likelihood estimation method (PIE) to address the challenge of information bias caused by phenotyping errors without specifying fixed values for the sensitivity and specificity of phenotyping algorithms. The proposed method utilizes the prior knowledge on the sensitivity and specificity through the integrated likelihood (IL) where the uncertainty on sensitivity and specificity is accounted for by integration. Such a method can mitigate the need for validation data, and can reduce bias in estimation of association due to fixing sensitivity and specificity at particular values. With simulation studies and real data examples, we demonstrate the advantage of this proposed method over the existing methods.

Missing at random or not: a semiparametric testing approach via instrumental variable

Rui Duan1, Jason Liang2, Pamela Shaw1, Chengyong Tang3 and Yong Chen1
1University of Pennsylvania  
2National Institute of Allergy and Infectious Disease  
3Temple University

Abstract: Missing at random or not: a semiparametric testing approach via instrumental variable

♦ Recalibrating Prognostic Risk Score Adapted to EHR Data

Dandan Liu, Qingxia Chen, Hui Nian and Joshua Denny
Vanderbilt University
dandan.liu@vanderbilt.edu

Prognostic risk score combine patient risk factors information into an accurate risk score and is valuable to clinicians advising patients, as well as policy makers and epidemiologists’ interest in risk adjustment. The widely adopted electronic health record (EHR) systems boost unique opportunities to automate implementation of risk prediction models into clinical decision support tools and conduct EHR-based health policy research using existing risk scores as risk adjustment. However, when researchers carried an existing prognostic risk score to EHR systems, problems often arose when some risk factor information are not collected in the exact form required for the risk calculation and only coarsely available in EHR. The conventional approach omitting such risk factors from the risk score calculation will result in biased risk prediction and risk adjustment and thus is not recommended. Statistical methods appropriately replace uncollected risk factor with coarsely available risk factor information in the existing risk scores are warranted. In this paper, we evaluated several methods to optimally combine such coarsely collected risk factors into existing risk scores. Our methods are based on minimization of several loss functions based on likelihood function, area under the curve (AUC), and discrimination slope (DSL). Simulation studies were conducted under various scenario to compare the proposed methods with the commonly used omitting approach. The proposed approaches were applied to adapt automated implementation of a prognostic mortality index in EHR.

Session 85: Longitudinal, Multilevel, Multiway and Spatial Functional Data Analysis

Modeling continuous glucose monitoring (CGM) data during sleep

Ciprian Crainiceanu
Johns Hopkins University
craini1@jhu.edu

We introduce a multilevel functional Beta model to quantify the blood glucose levels measured with continuous glucose monitors (CGMs) for multiple days in study participants with Type 2 diabetes mellitus. We focus on the actigraphy-estimated sleep periods to reduce the effects of meals and physical activity. The model produces confidence intervals for blood glucose levels at any time from the actigraphy-estimated sleep onset, quantifies the within- and between-subject variability of blood glucose, and produces interpretable parameters of the blood dynamics during the actigraphy-estimated sleep periods. We provide visualization tools for our results and validate the estimated model parameters versus levels of Hemoglobin A1c.

Dynamic Prediction of Alzheimer’s Disease Using Multiple Longitudinal Outcomes and Survival Data

Kan Li1 and Sheng Luo2
1Merck Research Laboratory, Merck & Co.  
2Duke University
sheng.luo@duke.edu

This paper is motivated by combining serial neurocognitive assessments and other clinical variables for monitoring the progression of Alzheimer’s disease (AD). We propose a novel framework for the use of multiple longitudinal neurocognitive markers to predict the progression of AD. The conventional joint modeling longitudinal and survival data approach is not applicable when there is a
large number of longitudinal outcomes. We introduce various approaches based on functional principal component (FPC) for dimension reduction and feature extraction from multiple longitudinal outcomes. We use these features to extrapolate the health outcome trajectories and use scores on these features as predictors in a Cox proportional hazards model to conduct predictions over time. We propose a personalized dynamic prediction framework that can be updated as new observations collected to reflect the patient’s latest prognosis, and thus intervention could be initiated in a timely manner. Simulation studies and application to the Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset demonstrate the robustness of the method for the prediction of future health outcomes and risks of target events under various scenarios.

Tests of significance for time-varying covariate effect in longitudinal functional data
*Ana-Maria Staicu*¹, Saebina Oh² and Arnab Maity¹
¹Carolina State University
²North Carolina State University
astaicu@ncsu.edu

We consider a time-varying functional regression model to describe association between one-dimensional curve responses and scalar covariates that are observed in a longitudinal design for many subjects. We develop a hypothesis testing procedure to assess the significance of the time-varying covariate effect. We propose a pseudo generalized F testing procedure that accounts for the complex error structure and is computationally efficient. Numerical studies confirm that the testing approach has the correct size and compares favorably with available competitors in terms of power. The methods are illustrated on a data application.

Covariance Function Estimation for Multidimensional Functional Data
Jiayi Wang¹, Raymond K. W. Wong³ and Xiaoke Zhang²
¹Texas A&M University
²George Washington University
raywong@tamu.edu

Multidimensional functional data are becoming more common in various domains such as climate studies, neuroimaging and chemometrics. In this talk I will present a nonparametric covariance function estimator for multidimensional functional data. It is based on an efficient spectral regularizations of an operator associated with a reproducing kernel Hilbert space. I will discuss the corresponding numerical and theoretical results.

Session 86: Random forests and decision trees for survival data

Subgroup Identification using Covariate Adjusted Interaction Trees
*Jon Steingrimsson* and *Jiabei Yang*
Brown University
jon_steingrimsson@brown.edu

We consider the problem of identifying sub-groups of participants in a clinical trial that have enhanced treatment effect. Recursive partitioning methods that recursively partition the covariate space based on some measure of between groups treatment effect difference are popular for such sub-group identification. The most commonly used recursive partitioning method, the classification and regression tree algorithm, first creates a large tree by recursively partitioning the covariate space using some splitting criteria and then selects the final tree from all subtrees of the large tree. In the context of subgroup identification, calculation of the splitting criteria and the evaluation measure used for final tree selection rely on comparing differences in means between the treatment and control arm. When covariates are prognostic for the outcome, covariate adjusted estimators have the ability to improve efficiency compared to using differences in means between the treatment and control group. We discuss properties and final sample performance of two covariate adjusted estimators that can be used to both make splitting decisions and for final tree selection.

Oblique Random Survival Forest
*Byron Jaeger¹, Leann Long¹, Dustin Long¹, Leslie Mcclure², Jeff Szczodrski¹, George Howard¹, Noah Simon, Yuan-I Min³ and Mario Sims³
¹University of Alabama at Birmingham
²Drexel University
³University of Mississippi Medical Center
bcjaeger@uab.edu

We introduce and evaluate the oblique random survival forest (ORSF). The ORSF is an ensemble method for right-censored survival data that uses linear combinations of input variables to recursively partition a set of training data. Regularized Cox proportional hazard models are used to identify linear combinations of input variables in each recursive partitioning step. Benchmark results using simulated and real data indicate that the ORSF’s predicted risk function has high prognostic value in comparison to random survival forests, conditional inference forests, regularized regression, and boosting. In an application to data from the Jackson Heart Study, we demonstrate variable and partial dependence using the ORSF and highlight characteristics of its 10-year predicted risk function for atherosclerotic cardiovascular disease events (ASCVD; stroke, coronary heart disease). We compare variable and partial effects of the ORSF with the CIF to highlight areas of agreement and discordance between the two ensemble methods. All computations with ORSF are based on the obliqueRF R package, available on the comprehensive R archive network (CRAN).

Treatment decision using causal survival forests
*Yifan Cui¹, Michael Kosorok², Stefan Wager³ and Ruoning Zhu³
¹University of Pennsylvania
²University of North Carolina at Chapel Hill
³Stanford University
³University of Illinois at Urbana-Champaign
yifan_cui@wharton.upenn.edu

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 data, estimate treatment effects and final sample performance of two covariate adjusted estimators that can be used to both make splitting decisions and for final tree selection.

Penalized Random Survival Forests
*Sarah Formentini* and *Ruoning Zhu*
University of Illinois at Urbana-Champaign
sarahf2@illinois.edu

Random survival forests are popular statistical models in biomedical studies, especially for cancer studies with high-dimensional genetic information. With the abundance of cancer genetics and genomics...
data, new studies can borrow information from existing ones. For this purpose, we propose penalized random survival forests that utilizes information from existing data to improve the model fitting. The penalization is achieved by constructing a new type of splitting rule that shrinks the marginal scores of a potential split. This new split has the potential to improve the convergence rate of random forest models. We experimented with two types of shrinkage methods by utilizing two types of existing information: the marginal p-value or summary statistics which are often released from existing studies, or the variable importance measure calculated from the existing data if the complete data are available. We perform extensive simulation studies to demonstrate the superior performance over existing single data set approaches, and also apply our method to genetic data for analyzing cancer.

Session 87: New developments in nonparametric statistics and empirical likelihood

Empirical likelihood inference for the panel count data with informative observation process  
Faysal Satter  
Georgia State University  
fsatter1@student.gsu.edu

Panel count data refer to interval-censored recurrent event data, which often arise from longitudinal studies. Each study subject can only be observed at discrete time points rather than continuously. As a result, one can only know the total number of events that occurred between two observation time points instead of the actual time of the events. Furthermore, the observation times can be different among subjects and carry important information about the underlying recurrent process. In this paper, an empirical likelihood (EL) method for panel count data with informative observation times is proposed. An empirical likelihood ratio for the vector of regression coefficients is formulated and the Wilks theorem is established. Simulation studies are carried out to show the performance of empirical likelihood and to compare those with normal approximation methods. We compare the EL with existing methods using an illustrative example from a bladder cancer study.

Collaborative Spectral Clustering in Attributed Networks  
Pengsheng Ji  
University of Georgia  
pjil@uga.edu

We proposed a novel spectral clustering algorithm for attributed networks, where n nodes split into R non-overlapping communities and each node has a p-dimensional meta covariate from various formats such as text, image, speech etc.. The connectivity matrix $W_{n \times n}$ is constructed with the adjacent matrix $A_{n \times n}$ and covariate matrix $X_{n \times p}$, and $W = (1 - \alpha)A + \alpha K(X, X')$, where $\alpha \in [0, 1]$ is a tuning parameter and $K$ is a Kernel to measure the covariate similarities. We then perform the classical k-means algorithm on the element-wise ratio matrix of the first K leading eigenvector of $W$. Theoretical and simulation studies showed the consistent performance under both Stochastic Block Model (SBM) and Degree-Corrected Block Model (DCBM), especially in imbalanced networks where most community detection algorithms fail.

Empirical Likelihood Inference for Generalized Additive Partially Linear Models  
\textsuperscript{1}Rong Liu and \textsuperscript{2}Yichuan Zhao  
\textsuperscript{1}University of Toledo  
\textsuperscript{2}Georgia State University  
rong.liu@utoledo.edu

Generalized additive partially linear models enjoy the simplicity of generalized linear models and the flexibility of generalized additive models because they combine both parametric and nonparametric components. Based on spline-backfitted kernel (SBK) estimator, we propose empirical likelihood (EL) based pointwise confidence intervals and simultaneous confidence bands (SCBs) for the non-parametric component functions to make statistical inference. Simulation study strongly supports the asymptotic theory, and shows that EL based SCBs are much easier for implementation and have better performance than Wald-type SCBs. We apply the proposed method to a university retention study and provide SCBs for the effect of the students information.

Jackknife Empirical Likelihood Approach for K-sample Tests via Energy Distance  
\textsuperscript{1}Yongli Sang, \textsuperscript{2}Xin Dang and \textsuperscript{3}Yichuan Zhao  
\textsuperscript{1}University of Louisiana at Lafayette  
\textsuperscript{2}University of Mississippi  
\textsuperscript{3}Georgia State University  
yongli.sang@louisiana.edu

Energy distance is a statistical distance between the distributions of random variables, which characterizes the equality of the distributions. Utilizing the energy distance, we develop a nonparametric test for the equality of K (K > 2) distributions in this talk. By applying the jackknife empirical likelihood approach, the standard limiting chi-square distribution with degrees freedom of K - 1 is established and is used to determine critical value and p-value of the test. Simulation studies show that our method is competitive to existing methods in terms of power of the tests in most cases. The proposed method is illustrated in an application on a real data set.

Session 88: Statistical Controversies in Forensic Evidence Interpretation

The Critical Role of Statistics in Evaluating Forensic Evidence  
Karen Kafadar  
University of Virginia  
kkafadar@virginia.edu

Statisticians have been valuable contributors to many areas of science, including chemistry (chemometrics), biology (genomics), medicine (clinical trials), and agriculture (crop yield), leading to important advances in statistical research that have benefitted multiple fields (e.g., spectral analysis, penalized regression, sequential analysis, experimental design). Statistics in forensic science has not been nearly as extensive, given its importance (ensuring proper administration of justice) and value it has demonstrated thus far (e.g., forensic DNA, assessment of bullet lead evidence, U.S. anthrax investigations, reliability of eyewitness identification). Forensic methods in many areas remain unvalidated, as recent investigations have highlighted (notably, bite marks and hair analysis). I will provide examples where statistics played a vital role in evaluating forensic evidence and suggest further areas where statisticians can make a big impact, towards the ultimate goal of strengthening forensic evidence to achieve its mission: to raise the level of confidence in the reliability of evidence in the criminal justice system.

Statistical Approaches for Analyzing and Presenting Forensic Footwear Impression Evidence  
Steven Lund  
National Institute of Standards and Technology  
steven.lund@nist.gov
Footwear impressions are often found at crime scenes and can be useful evidence in linking a shoe belonging to a person of interest to the scene of the crime. Trained forensic footwear examiners compare crime scene impressions with test impressions obtained from a shoe of interest and assess the probative value of this evidence. Many researchers have explored quantitative methods for footwear impressions and for quantifying the value of evidence. Rather than emphasizing the probative value interpretation provided by an expert or algorithm, we have sought to effectively summarize and present the empirical basis an audience may use to arrive at their own assessment of probative value. The NIST forensic footwear research team has developed an end-to-end system for this purpose using a hybrid of human perception and algorithmic comparison. This talk will summarize our research and key findings.

Which statistical paradigm should I use for forensic evidence interpretation?

♦ Danica Ommen1 and Christopher Saunders2

1 Iowa State University
2 South Dakota State University

dommen@iastate.edu

Traditionally, there were two different methods of interpreting forensic evidence that were used in practice: the “two-stage approach” and the “likelihood ratio approach.” The two-stage approach in forensic science was developed by Parker in 1966 and uses concepts from the classical paradigm of statistics, like traditional hypothesis testing with p-values and rejection regions. In contrast, the likelihood ratio approach uses concepts from the Bayesian paradigm of statistics, like Bayes Factors for model selection, and was developed for forensic science by Lindley in 1977. More recently, a new class of methods for forensic evidence interpretation is surfacing to deal with some of the issues encountered from using the traditional methods. These approaches are associated with the likelihood paradigm of statistics developed by Royall in 1997, and manifests in the form of score-based likelihood ratios and other attempts to approximate Bayes Factors for quantifying the value of forensic evidence. In this presentation, practical and philosophical advantages and disadvantages of each paradigm will be discussed. In addition, recent controversies and advances in research surrounding each of the paradigms will be explored.

Are reported likelihood ratios well calibrated?

♦ Jan Hannig1 and Hari Iyer2

1 University of North Carolina at Chapel Hill
2 National Institute of Standards and Technology
jan.hannig@unc.edu

Many machine learning algorithms used in forensic practice have as their output Bayes factor sometimes called likelihood ratios. For example, it is not unusual to see a report that claims the DNA found at the crime scene is 1,000,000 times more likely under the assumption that the defendant is the source than under the assumption that someone other than the defendant is the source. In this talk we summarize existing approaches for examining the validity of likelihood ratio systems and discuss a new statistical methodology based on generalized fiducial inference for empirically examining the validity of such likelihood ratio claims.

Session 89: Latent attribute models and their applications

Structured Latent Class Regression Analysis of Multivariate Binary Data: Estimating Disease Etiology

Zhenke Wu
University of Michigan
zhenkewu@umich.edu

Unavailability of gold-standard measurements presents a major barrier to epidemiology studies of disease etiology. For example, Pneumonia Etiology Research for Child Health study (PERCH, O’Brien et al. 2019. The Lancet) measured among cases and controls the presence or absence of multiple pathogen targets in sites peripheral to the lung such as nasal cavity to estimate the distribution of causes of pneumonia and assign individual probabilities of causes. Inference of the population and individual disease etiologies has been addressed by nested partially-latent class models (Wu et al. 2016. Biostatistics). These models represent the distribution of each case’s observations as a mixture of components, each one representing a cause; The models also assume known classes among a subset of subjects, e.g., controls without lung infection. However, these models do not describe the effects of explanatory variables on disease etiology and on the measurement process. Further the model misfits a few pairwise covariations observed in the data which may be driven by the variation in key covariates such as season. To close this gap, we propose a family of latent class regression models for estimating disease etiology fractions based on multivariate binary data obtained in case-control studies. We let the case mixing weights vary with continuous and discrete covariates. We also allow the distribution of control measurements vary with covariates in a separate latent variable regression via multinomial logistic additive models, for which we propose a novel prior to encourage simpler functional forms (covariate-independent and fewer latent subclasses among controls whenever adequate). We demonstrate the superior performance of the proposed model via simulations. We apply the model to the motivating PERCH study to characterize the dependence of pneumonia etiology on season, age, HIV status and disease severity.

Estimation of Q-matrix with unknown number of attributes

♦ Yinghan Chen1, Ying Liu2, Steven Culpepper2 and Yaguo Chen2

1 University of Nevada-Reno
2 University of Illinois at Urbana-Champaign
yinghanc@unr.edu

Cognitive diagnosis models (CDMs) are widely used for providing fine-grained classification of multidimensional collection of discrete attributes. The application of CDMs is based on the specification of attribute requirement on educational tasks in what is known as the Q matrix. Existing methods must pre-specify the number of attributes in order to estimate Q. A misspecified Q might bias parameters and diagnostic classifications, so it is fundamental to develop methods to estimate the Q matrix. We present a Bayesian strategy for jointly estimating the number of attributes (the number of columns in Q) and the elements of Q for the deterministic inputs, noisy and gate (DINA) model. We propose the crimp-sampler to transit between matrices with different number of columns and estimate the underlying Q matrix and model parameters with a Gibbs sampler.

Learning Attribute Patterns in High-Dimensional Structured Latent Attribute Models

♦ Yuqi Gu and Gongjun Xu

University of Michigan
yuqigu@umich.edu

Structured Latent Class Regression Analysis of Multivariate Binary Data: Estimating Disease Etiology

Zhenke Wu
University of Michigan
zhenkewu@umich.edu

Unavailability of gold-standard measurements presents a major barrier to epidemiology studies of disease etiology. For example, Pneumonia Etiology Research for Child Health study (PERCH, O’Brien et al. 2019. The Lancet) measured among cases and controls the presence or absence of multiple pathogen targets in sites peripheral to the lung such as nasal cavity to estimate the distribution of causes of pneumonia and assign individual probabilities of causes. Inference of the population and individual disease etiologies has been addressed by nested partially-latent class models (Wu et al. 2016. Biostatistics). These models represent the distribution of each case’s observations as a mixture of components, each one representing a cause; The models also assume known classes among a subset of subjects, e.g., controls without lung infection. However, these models do not describe the effects of explanatory variables on disease etiology and on the measurement process. Further the model misfits a few pairwise covariations observed in the data which may be driven by the variation in key covariates such as season. To close this gap, we propose a family of latent class regression models for estimating disease etiology fractions based on multivariate binary data obtained in case-control studies. We let the case mixing weights vary with continuous and discrete covariates. We also allow the distribution of control measurements vary with covariates in a separate latent variable regression via multinomial logistic additive models, for which we propose a novel prior to encourage simpler functional forms (covariate-independent and fewer latent subclasses among controls whenever adequate). We demonstrate the superior performance of the proposed model via simulations. We apply the model to the motivating PERCH study to characterize the dependence of pneumonia etiology on season, age, HIV status and disease severity.

Estimation of Q-matrix with unknown number of attributes

♦ Yinghan Chen1, Ying Liu2, Steven Culpepper2 and Yaguo Chen2

1 University of Nevada-Reno
2 University of Illinois at Urbana-Champaign
yinghanc@unr.edu

Cognitive diagnosis models (CDMs) are widely used for providing fine-grained classification of multidimensional collection of discrete attributes. The application of CDMs is based on the specification of attribute requirement on educational tasks in what is known as the Q matrix. Existing methods must pre-specify the number of attributes in order to estimate Q. A misspecified Q might bias parameters and diagnostic classifications, so it is fundamental to develop methods to estimate the Q matrix. We present a Bayesian strategy for jointly estimating the number of attributes (the number of columns in Q) and the elements of Q for the deterministic inputs, noisy and gate (DINA) model. We propose the crimp-sampler to transit between matrices with different number of columns and estimate the underlying Q matrix and model parameters with a Gibbs sampler.

Learning Attribute Patterns in High-Dimensional Structured Latent Attribute Models

♦ Yuqi Gu and Gongjun Xu

University of Michigan
yuqigu@umich.edu
Structured latent attribute models (SLAMs) are a special family of discrete latent variable models widely used in social and biological sciences. This paper considers the problem of learning significant attribute patterns from a SLAM with potentially high-dimensional configurations of the latent attributes. We address the theoretical identifiability issue, propose a penalized likelihood method for the selection of the attribute patterns, and further establish the selection consistency in such an overfit SLAM with diverging number of latent patterns. The good performance of the proposed methodology is illustrated by simulation studies and two real datasets in educational assessment.

#### Identified Bayesian Factor Analysis

*Albert Man and Steven Culpepper*

University of Illinois at Urbana-Champaign

[aman3@illinois.edu](mailto:aman3@illinois.edu)

Exploratory factor analysis is a dimension-reduction technique commonly used in psychology, finance, and economics. Advances in computational power have opened the door for fully Bayesian treatments of factor analysis. One open problem is enforcing identifiability of the latent factor loadings, as the loadings are not identified from the likelihood without further restrictions. Nonidentifiability of the loadings can cause posterior multimodality, which can produce misleading posterior summaries. The positive-diagonal, lower-triangular (PLT) constraint is the most commonly used restriction to guarantee identifiability, in which the upper $m \times m$ submatrix of the loadings is constrained to be a lower-triangular matrix with positive-diagonal elements. The PLT constraint can fail to guarantee identifiability if the constrained submatrix is singular. Additionally, we will show that though the PLT constraint can resolve identifiability-related multimodality, it will introduce additional mixing issues. We propose to treat the row indices of the constrained PLT submatrix as a random variable, allowing the PLT structure to be embedded in any subset of rows of the loadings matrix. We then develop a mode-jumping metropolis-hastings step to (1) sample new row indices for the PLT structure and (2) rotate to new loadings under said constraints. We demonstrate the proposed algorithm’s performance in parameter estimation, mixing properties, and recovery of the correct number of factors.

#### Session 90: Recent Development in High-Dimensional Data Analysis

#### A two sample test of equality of means in high dimensional data

*Huaiyu Zhang and Haiyan Wang*

Kansas State University

[hwang@ksu.edu](mailto:hwang@ksu.edu)

This research is interested in testing equality of two sample means in high-dimensional data in which the sample sizes may be much less than the dimension. When the dimension is larger than the sample size, it is impossible to construct a uniformly most powerful test. Effort in recent literature modifies the Hotelling’s $T^2$-statistic by either bypassing the estimation of high-dimensional covariance matrices or estimating the precision matrix after imposing sparseness condition. I will present a new test statistic that improves the scaling parameter of the average squared component-wise $t$-statistic in GCT test (Gregory et al. 2015) by taking advantage of the multiple copies of the underlying dynamics. Our new statistic overcomes several limitations of the GCT test: (i) it removes the stationarity requirement implicitly used in the autocovariance estimation in the GCT test; (ii) the new estimator of scaling parameter is more efficient than that used in the GCT statistic. (iii) the test works well even when the component indices of the data vector have high correlations as long as such correlation reduces suitably fast as the separation of the component indices increases (at polynomial rate or faster). The limiting distribution of the test statistic and power of the test are studied. Simulation results and a real application will be presented to show the numerical performance of the test and to compare with other tests in the literature.

#### High-Dimensional Rank-Based Inference

*Solomon Harrar$^1$ and Xiaoli Kong$^2$*

$^1$University of Kentucky

[soharrar@uky.edu](mailto:soharrar@uky.edu)

Existing high-dimensional inferential methods for comparing multiple groups formulate hypothesis in terms of mean vectors or location parameters. These methods are applicable mainly for data where differences between values make sense. Furthermore, the mean-based methods assume existence of moments and the non-parametric (location-based) ones assume elliptical-contoured distributions for the populations. In this paper, a novel fully nonparametric (rank-based) method is proposed. The method is applicable for both metric and non-metric data and, hence, is applicable for ordered categorical as well as skewed and heavy tailed data. To develop the theory, we prove a novel result for studying asymptotic behavior of quadratic forms in ranks. Simulation study shows that the developed rank-based method performs comparably well with mean-based methods when the assumptions of those methods are satisfied. However, it has significantly superior power for heavy tailed distribution with the possibility of outliers. The rank method is applied to an EEG data with the objective of examining association between alcohol use and change in brain function.

#### Sufficient dimension reduction using a divergence measure for high dimensional data

*Qingcong Yuan$^1$, Wenhui Sheng$^2$ and Xiangrong Yin$^3$*

$^1$Miami University

$^2$Marquette University

$^3$University of Kentucky

[qingcong.yuan@miamioh.edu](mailto:qingcong.yuan@miamioh.edu)

Using expectation of conditional difference measure, we introduce a novel approach to sufficient dimension reduction problems. The proposed method is model-free and is especially useful when the response is categorical. The estimation of dimension and the central subspace using the measure is discussed and the dimension reduction for large $p$ and small $n$ cases is developed. The root-$n$ consistency and asymptotic normality is established under regularity conditions. Numerical studies are provided to demonstrate the advantage of the method. The proposed method is very competitive and robust compared to existing dimension reduction methods.

#### On sparse Fourier transform inverse regression for sufficient variable selection

*Jiaying Weng and Xiangrong Yin*

University of Kentucky

[jiaying.weng@uky.edu](mailto:jiaying.weng@uky.edu)

The big data era poses great challenges as well as opportunities for researchers to develop efficient statistical approaches to analyze massive data. Sufficient dimension reduction (SDR) is such an important tool in modern data analysis and has received extensive attention in both academia and industry. I introduce an inverse regression estimator using Fourier transforms, which is superior to the existing SDR methods in two folds, (1) it avoids the slicing of...
the response variable, (2) it can be readily extended to solve high dimensional data problem. For the ultra-high dimensional problem, I investigate a minimum discrepancy approach to achieve optimal solutions and also develop a novel and efficient optimization algorithm to obtain the sparse estimates. Simulation studies and real data analysis are carried out to illustrate the superior performance of our proposed methods.

Session 91: Novel genetic discovery and risk prediction methods for Alzheimer disease

Family-based association tests for rare variants with censored traits

• Wenjing Qi¹, Shuming Sun¹, X. Raymond Gao², Eden R. Martin³, Andrew S. Allen¹ and Yi-Ju Li¹

¹Duke University
²Ohio State University
³University of Miami

We propose a set of family-based burden and kernel tests for censored traits (FamBAC and FamKAC). Here, censored traits refer to time-to-event outcomes, for instance, age-at-onset of a disease. Testing for censored traits allow us to include not only affected subjects but also unaffected subjects, and it focuses on the relative risk (hazard ratio) of developing the disease at any age between the risk allele carriers and the non-carriers for the gene of interest. To model censored traits in family-based designs, we used the frailty model, which incorporated not only fixed genetic effects of rare variants in a region of interest but also random polygenic effects shared within families. We first partitioned genotype scores of rare variants into orthogonal between- and within-family components, and then derived their corresponding efficient score statistics from the frailty model. Finally, FamBAC and FamKAC were constructed by aggregating the weighted efficient scores of the within-family components across rare variants and subjects. FamBAC collapsed rare variants within subject first to form a burden test that followed a chi-squared distribution; whereas FamKAC was a variant component test following a mixture of chi-squared distributions. For FamKAC, p-values can be computed by permutation tests or for computational efficiency by approximation methods. Through simulation studies, we showed that type I error was correctly controlled by FamBAC for various variant weighting schemes. However, FamKAC type I error rates based on approximation methods were deflated but improved by permutation tests. Our simulations also demonstrated that burden test FamBAC had higher power than kernel test FamKAC when high proportion of causal variants had effects in the same direction. In contrast, when the effects of causal variants on the censored trait were in mixed directions, FamKAC outperformed FamBAC and had comparable or higher power than an existing method, RVFam. Our proposed framework has the flexibility to accommodate general nuclear families for rare variant analysis. We illustrate the application of our methods using nuclear family data from Alzheimer’s Disease Genetics Consortium (ADGC) Genome Wide Association Study - NIA Alzheimer’s Disease Centers Cohort. The development of an R package is underway.

Precise modelling zero-inflated count phenotypes from sequenc- ing data

• Qiao Fan¹, Shuming Sun² and Yi-Ju Li²

¹Centre for Quantitative Medicine, Duke-NUS
²Duke University Medical Center

Although phenotypes as continuous variable with normal distribution or binary outcome are widely seen in the genetic study, it is not uncommon to encounter other form of counting variable with excessive zero, such as neuritic plaques (NPs) in brain pathology studies in Alzheimer’s, small RNA-seq reads, etc. The outcome is a mixture distribution with one component being the “structural zero” and the other component being a Poisson distribution. Also, with the reduced cost in sequencing technology, more researchers tend to use NGS data to survey low and rare genetic variant. To assess the sequencing variants with excessive count data, we propose burden and kernel test by constructing efficient scores to test the association between rare variants and structured zeros, as well as rare variants and counts. We present the procedure of our proposed test statistics and evaluate its performance.

Post-GWAS data integration identifies risk factors for Alzheimer’s disease

Qiongshi Lu

University of Wisconsin-Madison

Despite the findings in genome-wide association studies (GWAS) for late-onset Alzheimer’s disease (LOAD), our understanding of its genetic architecture is far from complete. Transcriptome-wide association analysis that integrates GWAS data with large-scale transcriptomic databases is a powerful method to study the genetic architecture of complex traits. However, it is challenging to effectively utilize transcriptomic information given limited and unbalanced sample sizes in different tissues. Here we introduce and apply UTMOST, a principled framework to jointly impute gene expression across multiple tissues and perform cross-tissue gene-level association analysis using GWAS summary statistics. Compared with single-tissue methods, UTMOST achieved 39% improvement in expression imputation accuracy and generated effective imputation models for 120% more genes in each tissue. A total of 69 genes reached the Bonferroni-corrected significance level in the transcriptome-wide association meta-analysis for LOAD. Among these findings, we identified novel risk genes at known LOAD-associated loci as well as five novel risk loci. Several genes, including IL10 and ADRA1A, also have therapeutic potential to improve neurodegeneration. Cross-tissue conditional analysis further fine-mapped IL10 as the functional gene at the CR1 locus, a well-replicated risk locus for LOAD. Extension of this framework to perform biobank-wide association scans will also be discussed. Overall, integrated analysis of transcriptomic annotations and biobank information provides insights into the genetic basis of LOAD and may guide functional studies in the future.

Evaluation of polygenic risk scores for prediction of complex disease

• X. Raymond Gao¹, Yi-Ju Li² and Eden Martin³

¹Ohio State University
²Duke University
³University of Miami

Polygenic risk scores (PRSs) have proven valuable for predicting quantitative traits and disease risk in recent years. PRSs aggregate many individual variants into a single measure, typically weighted by their effect sizes. Suppose k independent single nucleotide polymorphisms (SNPs) are obtained from the association analyses, weighted PRS is defined as $\text{PRS} = \sum_{i=1}^{k} \hat{\beta}_i G_i$, where $\hat{\beta}_i$ and $G_i$ are the effect size from a regression analysis and the allele dosage for
each SNP $i$, respectively. Despite its calculation is straightforward, there is no consensus on how to select the best set of SNPs for the optimal prediction accuracy. A popular approach is to select SNPs that meet genome-wide significance threshold with $P < 5 \times 10^{-8}$. Sometimes, $P < 5 \times 10^{-5}$ is also used for selecting SNPs. In recent years, a number of methods used Bayesian methods to model the proportion of SNPs that shows non-zero effect sizes, such as the LDpred method, which eliminated the step of preselecting SNPs based on an arbitrary threshold, e.g. $5 \times 10^{-8}$. However, LDpred may not give the best prediction accuracy either in some situations and it can be memory demanding and time consuming. In this study, we evaluate a number of approaches to derive PRSs for both quantitative traits and disease risk using empirical and simulated datasets. Our results provide guidelines for performing PRSs for prediction of complex disease.

Session 92: Bayesian methodology and applications in clinical trials

A simple approach to incorporating historical data in clinical trial design and analysis

† Lanju Zhang, Zailong Wang, Li Wang, Lu Cui, Jeremy Sokolove and Ivan Chan

Abbvie, Inc.

Lanju.zhang@abbvie.com

Trials are not designed in a vacuum and historical control data are very often abundant. Use of available historical control data in a new trial can reduce the number of control patients and accordingly reduce both costs and timelines. On the other hand, bias, defined as the difference between historical and concurrent control data, can increase the false decision rate and confound the outcome interpretation. In this talk, we present a simple approach to incorporating historical control data in clinical trial design and analysis. We also propose a six-step process for trial design including selection and summarization of historical control data, sample size determination with and without borrowing, design property evaluation, and how much historical data should be borrowed to control false decision rate inflation based on trial variability. Examples are used to demonstrate the methodology and process.

Bayesian logistic mixed effects model for multiple imputation of frequently repeated binary outcomes

† Tianmeng Lyu, Björn Holzhauer, Dong Xi and Alkaz Uddin

Novartis Pharmaceuticals Corporation

tianmengpj@gmail.com

Treatment adherence is essential to ensure the clinical benefit of therapy. Good adherence to inhaled medication correlates with a lower risk of exacerbations for patients with COPD (Chronic Obstructive Pulmonary Disease). In practice, however, the adherence rate may be as low as 50% and decrease with time, which may result in a higher risk of exacerbations and hospitalization. The aim of a randomized controlled trial is to test whether patients using an adherence app will have improved adherence. During the follow-up, binary inhalation data will be recorded daily but may be missing due to drop-out or inhaler loss. Thus, an imputation method that targets the chosen hypothetical estimand is needed. We propose to perform multiple imputation by modeling the adherence in different time periods with a Bayesian logistic regression mixed effects model using vague priors. The model can account for the possible time trend in adherence. We implemented the proposed method in Stan for efficient sampling. Simulations were used to evaluate the performance of the proposed method and to compare it to the commonly used multivariate imputation by chained equations (MICE) approach. MICE imputes the missing binary data on each day sequentially using the covariates and both the observed and imputed data in previous days. The results show that our proposed method has improved performance over MICE.

Detecting Differentially Methylated Regions Accounting for Cell Purity Using Bayesian Credible Band

Shili Lin

Ohio State University

shili@stat.osu.edu

DNA methylation was the first epigenetic mechanism studied and remains the most investigated to date. One major goal of studying DNA methylation using data from high throughput assays such as bisulfite sequencing (BS-seq) or microarrays is to detect differentially methylated regions (DMRs) between two conditions of interest, such as diseased and normal. DMRs are of great biological relevance because they can provide strong evidence for association with gene expression and are frequently studied in cancer research. However, since solid tumor tissues not only contain tumor cells but also cells of other types (e.g. normal, stromal, and immune), DNA methylation from BS-seq or microarray is in fact a mixture from multiple cell types. As such, differential methylation signals between tumor cells and normal cells may be masked, leading to reduced power for detecting DMRs. In this talk, I will describe BCurve, a Bayesian credible bands method for detecting DMRs that take tumor cell purity into consideration. In addition, influential covariates, such as age, sex, and race, can also be accounted for. I will demonstrate the performance quality and utility of BCurve using simulated data as well as data from The Cancer Genome Atlas.

Compartment models using PROC NLMIXED

Raghavendra Rao Kurada

SAS Institute, Inc.

Raghavendra.kurada@sas.com

Pharmacokinetics (PK) is a branch of medical research that models the movement of a drug through the body. PK models are non-linear models that are widely used in the biopharmaceutical industry to predict pharmacokinetic changes in a body system. In SAS/STAT®14.3, the NLMIXED procedure provides enhanced PK modeling capability through the new CMPTMODEL statement, which enables you to fit a large class of PK models, including one-, two-, and three-compartment models for intravascular (bolus and infusion) and extravascular (oral) types of drug administration. The CMPTMODEL statement also supports PK models for multiple dosages and PK models with various parameterizations. In this talk, usage of this new statement is illustrated through a couple of examples.

Session 93: Statistical Methods for Integrative Analysis of Multi-Omics Data

Robust integrative data analysis using penalized weighting methods

Xiaoli Gao

University of North Carolina at Greensboro

xgao2@uncg.edu

With the development of high-throughput technologies, different types of omics data can be generated to provide a global view across tumors and normal samples. This data emergence has brought many
challenges to the statistical development for data integration. One of the most important challenges is the data contamination and heteroscedasticity. This paper studies a novel penalized weighting method to incorporate the outlier analysis into the data integration process, and therefore perform a robust Integrated Data Analysis and Biomarker Discovery. Simulation studies and real data analysis are implemented to demonstrate the performance of proposed approach.

An efficient prospective likelihood approach to secondary trait-genetic association analyses

Xi Lu and Cen Wu

University of Chicago

Gene environment interactions (GxE) play important roles in elucidating the disease etiology, in addition to genetic and environmental main effect. Traditionally, GXE studies have been pursued by considering a single platform G factor, such as the gene expressions and SNPs. As multi-level omics feature has been made available, incorporating the regulator information in GXE studies is expected to shed novel insight beyond the single level analysis. In this study, we develop an efficient two step robust penalization method to accommodate the integration of multi-dimensional omics measurements in GXE studies. In the first stage, we identify the sparse regulatory relationship, which has been accounted for when searching for important main and interaction effects associated with the cancer outcome robustly in the second stage. We demonstrate the advantage of the proposed method through extensive simulation studies. In the case study of TCGA lung cancer data, our method leads to findings with important implications.

Session 94: Innovative methods for handling missing data in the era of data science

Bayesian profiling multiple imputation for missing electronic health records

Yajuan Si, Mari Palta and Maureen Smith

University of Michigan

University of Wisconsin-Madison

Electronic health records (EHRs) are increasingly popular for clinical and comparative effectiveness research but suffer from usability deficiencies. Motivated by health services research on diabetes care, we seek to increase the quality of EHRs and focus on missing longitudinal glycosylated hemoglobin (A1C) values. Under the framework of multiple imputation we propose an individualized Bayesian latent profiling approach to capturing A1C measurement trajectories related to missingness. We examine different missingness mechanisms and perform model diagnostics and sensitivity analysis. The proposed method is applied to EHRs of adult patients with diabetes who were medically housed in a large academic Midwestern health system between 2003 and 2013. Our approach fits flexible models with computational efficiency and provides useful insights into the clinical setting.

A simple Index of Local Sensitivity to Non-Ignorability for Intensive Longitudinal Data with missing

Hui Xie, Chengbo Yuan, Donald Hedeker and Robin Mermelstein

Simon Fraser University

University of Illinois at Chicago

When analyzing intensive longitudinal data such as Ecological Momentary Assessment (EMA) data, it is often assumed the missingness is ignorable. Since this assumption is unverifiable, it is important to perform sensitivity analysis to assess the potential impact of non-ignorability. However a sensitivity analysis that directly fits a range of nonignorable models can be challenging to conduct because the resulting likelihood functions from these nonignorable models in intensive longitudinal data involve high dimensional integrations, due to missingness in both outcome and covariates as well as non-monotone missingness patterns in intensive longitudinal data. Gao et al. [2016] proposed a method using both linear and non-linear index of sensitivity to nonignorability (ISNIL and ISNIQ) to measure the local sensitivity of the missing at random (MAR) estimators to nonignorability for cross-sectional data. We extend the method to intensive longitudinal data. We considered both outcome and covariates to follow multivariate normal distribution situations and model the non-monotone missingness using transitional multinomial models. This method largely simplifies the computation process and substantially reduce computational workload. We derive the formulas for ISNI index measures and evaluate the performance of the extended method using simulated data, and apply it to a real EMA dataset.

Monte Carlo Approach to Likelihood Maximization for Missing Data Problems

Hua Yun Chen

University of Illinois at Chicago

When analyzing intensive longitudinal data such as Ecological Momentary Assessment (EMA) data, it is often assumed the missingness is ignorable. Since this assumption is unverifiable, it is important to perform sensitivity analysis to assess the potential impact of non-ignorability. However a sensitivity analysis that directly fits a range of nonignorable models can be challenging to conduct because the resulting likelihood functions from these nonignorable models in intensive longitudinal data involve high dimensional integrations, due to missingness in both outcome and covariates as well as non-monotone missingness patterns in intensive longitudinal data. Gao et al. [2016] proposed a method using both linear and non-linear index of sensitivity to nonignorability (ISNIL and ISNIQ) to measure the local sensitivity of the missing at random (MAR) estimators to nonignorability for cross-sectional data. We extend the method to intensive longitudinal data. We considered both outcome and covariates to follow multivariate normal distribution situations and model the non-monotone missingness using transitional multinomial models. This method largely simplifies the computation process and substantially reduce computational workload. We derive the formulas for ISNI index measures and evaluate the performance of the extended method using simulated data, and apply it to a real EMA dataset.
Maximizing the incomplete data likelihood can be challenging due to the intractable integral in obtaining the observed data likelihood. The EM algorithm is often used to reduce the complexity of the maximization and to increase the stability of the algorithm. However, the E-step in the EM algorithm can also have intractable integral. In such a situation, the Monte Carlo EM algorithm was proposed to approximate the conditional expectations in the E-step by Monte Carlo simulation. When Monte Carlo samples are available, the Monte Carlo EM algorithm is inefficient due to the inherent slow convergence rate of the EM algorithm. Direct likelihood maximization is more attractive. We propose a Monte Carlo approach to iteratively approximating the observed data likelihood and maximizing the approximate observed data likelihood to obtain the maximum likelihood estimator. The proposed approach can substantially reduce the amount of samples to be drawn due to its fast convergence speed. The iterative approximation and maximization approach for the likelihood maximization is also accurate enough so that the approximate likelihood can be directly used for inference such as the likelihood ratio test. We demonstrate the advantages of the proposed approach over the Monte Carlo EM algorithm by a number of examples.

Robust Estimation of the Average Treatment Effect under Data Combination
BaoLuo Sun
National University of Singapore
stasb@nus.edu.sg

Although instrumental variable methods are frequently used to estimate causal effects in the presence of unmeasured confounding, the exposure and outcome are often recorded in separate datasets due to complex data harvesting procedures. As a remedy, two-sample IV methods have been widely applied in the health and social sciences to elucidate novel causal mechanisms under data combination. We investigate issues of efficiency and robustness of existing estimators under a semiparametric framework for the two-sample IV problem, and propose multiply robust, locally efficient estimators of causal effects under a semiparametric framework for the two-sample IV problem, to elucidate novel causal mechanisms under data combination. We compared our approach with existing deconvolution based methods (e.g. DeMix) and demonstrated better performance in simulations studies. We applied this approach to Clinical Proteomic Tumor Analysis Consortium’s (CPTAC) clear cell renal cell carcinoma data for genomic, transcriptomic and proteomic integrative analyses in tumor.

A Unified Approach to Sparse Tweedie Modeling of Multi-Source Insurance Claim Data
Yaowen Gie1, Simon Fontaine2, Yi Yang3, Wei Qian4 and Bo Fan5
1University of Connecticut
2University of Montreal
3McGill University
4University of Delaware
5University of Oxford
knightqu34@gmail.com

Actuarial practitioners now have access to multiple sources of insurance data corresponding to various situations: multiple business lines, umbrella coverage, multiple hazards, and so on. Despite the wide use and simple nature of single-target approaches, modeling these types of data may benefit from a simultaneous approach. We propose a unified algorithm to perform sparse learning of such fused insurance data under the Tweedie (compound Poisson) model. By integrating ideas from multi-task sparse learning and sparse Tweedie modeling, our algorithm produces flexible regularization that balances predictor sparsity and between-sources sparsity. When applied to simulated and real data, our approach clearly outperforms single-target modeling in both prediction and selection accuracy, notably when the sources do not have exactly the same set of predictors. An efficient implementation of the proposed algorithm is provided in our R package MStweedie.

Identifiability of Restricted Latent Class Models
Gongjun Xu and Yuqi Gu
University of Michigan
gongjun@umich.edu

Latent class models have wide applications in social and biological sciences. In many applications, pre-specified restrictions are imposed on the parameter space of latent class models, through a design matrix, to reflect practitioners’ diagnostic assumptions about how the observed responses depend on the respondents’ latent traits. Though widely used in various fields, such restricted latent class models suffer from nonidentifiability due to the models’ discrete nature and complex restricted structure. This work addresses the fundamental identifiability issue of restricted latent class models by developing a general framework for strict and partial identifiability of the model parameters. The developed identifiability conditions only depend on the design matrix and are easily checkable, which provides useful practical guidelines for designing statistically valid diagnostic tests. Furthermore, the new theoretical framework is applied to establish, for the first time, identifiability of several designs from cognitive diagnosis applications.

Session 96: Recent Advances in Survival Methods for Clinical Trials

Biomarker-integrated Clinical Trials with Threshold Selection and Enrichment
Xiaofei Wang1, Ting Wang2 and Stephen George3
1Duke University Medical Center
2University of North Carolina at Chapel Hill
xiaofei.wang@duke.edu

A biomarker with potential to guide treatments is often measured on a continuous scale and optimal cutoffs for patient subgrouping are not available at the time of trial design. Biomarker-driven RCT designs have been proposed to simultaneously search for optimal cutoffs for subgroup identification and to evaluate treatment effects...
bivariate outcomes of toxicity/efficacy

Time-to-event continual reassessment method (TITE-CRM) for bivariate outcomes of toxicity/efficacy

Donglin Yan\textsuperscript{1}, Nolan Wages\textsuperscript{2}, Christopher Tait\textsuperscript{3}, Tamila Kindwall-Keller\textsuperscript{2} and Emily Dressler\textsuperscript{4}

\textsuperscript{1}Medpace
\textsuperscript{2}University of Virginia
\textsuperscript{3}PRA Health Sciences
\textsuperscript{4}Wake Forest University

dressle@wakehealth.edu

We consider the challenge of designing Phase I-II clinical trials with delayed toxicity and efficacy outcomes, as motivated by a Phase I-II study evaluating all-trans retinoic acid (ATRA) in combination with a fixed dose of daratumumab in the treatment of relapsed or refractory multiple myeloma. The primary objective of the study is to identify a dose that maximizes efficacy and has an acceptable level of toxicity. The toxicity endpoint is observed in one cycle of therapy (i.e., 4 weeks) while the efficacy endpoint is assessed after 8 weeks of treatment. The difference in endpoint observation windows causes logistical challenges in conducting the trial, since it is not practical to wait until both outcomes for each patient have been fully observed before sequentially assigning the dose of a newly eligible patient. In order to avoid delays in treatment for newly enrolled patients and to accelerate trial progress, we generalize the time-to-event continual reassessment method (TITE-CRM) to bivariate outcomes. Simulation studies evaluate the proposed method, and results show the proposed design is able to accurately select doses that maximize efficacy and have acceptable toxicity, while using all available information in allocating patients at the time of dose assignment. We compare the proposed methodology to two existing methods in the area.

Optimal Two-Stage Phase II Survival Trial Design

Jianrong Wu, Li Chen, Jing Wei and Heidi Weiss

University of Kentucky
jianrong.wu@uky.edu

Recently molecularly targeted agents and immunotherapy have been advanced for the treatment of relapse or refractory cancer patients, where disease progression-free survival or event-free survival is often a primary endpoint for the trial design. However, methods to evaluate two-stage single-arm phase II trials with a time-to-event endpoint are currently processed under an exponential distribution, which limits application of real trial designs. In this paper, we developed an optimal two-stage design, which is applied to the four commonly used parametric survival distributions and a non-parametric log spline distribution. The proposed method has advantages compared to existing methods in that the choice of underlying survival model is more flexible and the power of the study is more adequately addressed. Therefore, the proposed optimal two-stage design can be routinely used for single-arm phase II trial designs with a time-to-event endpoint as a complement to the commonly used Simon’s two-stage design for the binary outcome

Phase II trial design with growth modulation index as the primary endpoint

Jianrong Wu\textsuperscript{1}, Li Chen, Jing Wei, Heidi Weiss, Rachel Miller and John Villano

University of Kentucky
lichenky@uky.edu

Molecularly targeted, genomic-driven and immunotherapy-based clinical trials continue to be advanced for the treatment of relapse or refractory cancer patients, where the growth modulation index GMI is often considered a primary endpoint of treatment efficacy. However, there is little literature is available that consider the trial design with GMI as the primary endpoint. In this article, we derived a sample size formula for the score test under a log-linear model of the GMI. Study designs using the derived sample size formula are illustrated under a bivariate exponential model, the Weibull frailty model and the generalized treatment effect size. The proposed designs provide sound statistical methods for a single-arm phase II trial with GMI as the primary endpoint.

Session 97: New developments in big data science and biomedical applications

Detecting Dense and Sparse Signals in Biomedical Applications

Chi Song, Lo-Bin Chang and Xiaoyu Cai

Ohio State University
song.1188@osu.edu

In biomedical research, there is a common scenario when we are interested in testing whether there exists any signal of unknown proportion in a dataset that contains tremendous amount of noise. For example, in chromosome copy number detection problem, the goal is to detect the unknown proportion of carriers of certain copy number change among non-carriers; and in GWAS, we are interested in detecting an unknown proportion of SNPs in a SNP set that may associate with a certain disease. In this talk, I will propose a novel approach that adaptively combines statistical tests to detect both dense and sparse signals in high dimensional data. Simulation will be used to explore the properties of this approach around the theoretical detection boundary. I will also discuss the application of this new approach in DNA copy number detection and SNP set association test.

Multivariate Longitudinal Analysis with Functional Copula Models

Colin O. Wu

National Institutes of Health
wuc@nhlbi.nih.gov

A major objective of longitudinal studies is to evaluate the time-varying patterns of the conditional distribution functions for multivariate outcomes over time. Existing methods for longitudinal analysis focus on modeling the conditional means, correlations and distributions separately using different regression models. These approaches lack cohesiveness when the scientific objective requires a model that simultaneously describes the conditional means, correlations and distributions. We develop a class of nonparametric functional vine Copula models that incorporates the time-varying means, covariances and distributions into a regression structure. In this approach, we assume that the conditional distributions of the outcome variables belong to some well-known copula families at
fixed time points, while the copula parameters are functions of time. We propose a B-spline approximation method to estimate the functional parameters and distributions, and demonstrate its appropriate statistical properties through an epidemiological study of childhood cardiovascular risks and a simulation study. Theoretical justification for the B-spline approximation method is demonstrate through its consistency.

A unified machine learning approach to estimate the clinically meaningful changes
Jiwei Zhao
State University of New York at Buffalo
jiwei2012zhao@gmail.com

Understanding the limitation of solely using statistical significance, researchers have proposed to draw conclusions in biomedical studies based on clinical significance. The minimal clinically important significance (MCID) is one of the most fundamental concepts to study clinical significance. To estimate MCID based on an anchor question usually available in the patients’ reported outcomes, Hedaya, Wang and Xu (2015) presented a method using the classification technique. However, their method implicitly requires the probability that the anchor question has a positive outcome to be 0.5, a balanced outcome situation. This balancedness can not be controlled when one designs the study hence can not be guaranteed. In this paper, we propose a data adaptive method, which does not need a balanced outcome hence overcomes this restriction. Compared to Hedaya, Wang and Xu (2015), our method uses a faster gradient based algorithm and adopts a more flexible structure of the MCID at its individual level. We conduct comprehensive simulation studies and apply our method to the ChAMP study to demonstrate its usefulness and also its outperformance.

Session 98: New estimation, model selection and optimization for non-traditional data

IMPUTED FACTOR REGRESSION FOR HIGH-DIMENSIONAL BLOCK-WISE MISSING DATA

Yanqing Zhang\textsuperscript{1}, Annie Qu\textsuperscript{2} and Niansheng Tang\textsuperscript{1}
\textsuperscript{1}Yunnan University
\textsuperscript{2}University of Illinois at Urbana-Champaign
yanqingz@illinois.edu

Block-wise missing data arise more frequently nowadays in high-dimensional biomedical studies, social, psychological and environmental sciences, and there is an urgent need to develop efficient dimension reduction to extract important information for prediction under block-wise missing data. Existing dimension reduction methods and feature combination are ineffective for handling block-wise missing data. We propose the factor-model imputation approach targeting block-wise missing and model imputed factor regression for dimension reduction and prediction. Specifically, we first perform screening to identify the important features, impute important features based on the factor model, and then build a factor regression model to predict the response variable based on the important features imputed. The proposed method utilizes the essential information from all observed data through the factor structure model, and shows that it is still efficient even when the block-wise missing proportion is high. We show that the imputed factor regression model and its prediction are consistent under regularity conditions. We compare the proposed method with other existing approaches through simulation studies and a real data application to the Alzheimer’s Disease Neuroimaging Initiative (ADNI) data. Our numerical studies confirm that the proposed method outperforms the existing competitive approaches.

Integrating multi-source block-wise missing data in model selection

Fei Xue and Annie Qu
University of Illinois at Urbana-Champaign
feixue4@illinois.edu

For multi-source data, blocks of variable information from certain sources are likely missing. Existing methods for handling missing data do not take structures of block-wise missing data into consideration. In this paper, we propose a Multiple Block-wise Imputation (MBI) approach, which incorporates imputations based on both complete and incomplete observations. Specifically, for a given missing pattern group, the imputations in MBI incorporate more samples from groups with fewer observed variables in addition to the group with complete observations. We propose to construct estimating equations based on all available information, and optimally integrate informative estimating functions to achieve efficient estimators. We show that the proposed method has estimation and model selection consistency under both fixed-dimensional and high-dimensional settings. Moreover, the proposed estimator is asymptotically more efficient than the estimator based on a single imputation from complete observations only. In addition, the proposed method is not restricted to missing completely at random. Numerical studies and ADNI data application confirm that the proposed method outperforms existing variable selection methods under various missing mechanisms.

A Parsimonious Personalized Dose Finding Model via Dimension Reduction

Wenzhuo Zhou and Ruoqing Zhu
University of Illinois at Urbana-Champaign
wenzhuo3@illinois.edu

Learning an individualized dose rule (IDR) in personalized medicine is a challenging statistical problem. Existing methods for estimating the optimal IDR often suffer from the curse of dimensionality, especially when the IDR is learned nonparametrically using machine learning approaches. To tackle this problem, we propose a dimension reduction framework. The proposed methods exploit that the IDR can be reduced to a nonparametric function which relies only on a few linear combinations of the original covariates, hence leading to a more parsimonious model. To achieve this, we propose two approaches, a direct learning approach that yields the IDR as commonly desired in personalized medicine, and a pseudo-direct learning approach that focuses more on learning the dimension reduction space. Under regularity assumptions, we provide the convergence rate for the semiparametric estimators and Fisher consistency properties for the corresponding value function. For the pseudo-direct learning estimator, we use an orthogonality constrained optimization approach on Stiefel manifold to update the dimension reduction space. For the direct learning approach, we use an alternative updating scheme that iteratively updates the dimension reduction space and the nonparametric optimal dose rule function. The performances of the proposed methods are evaluated through simulation studies and a warfarin pharmacogenetic dataset.

Session 99: Recent Advances in Latent Variable Modeling

An Exploration of Latent Structure in Process Data
Xueying Tang, Zhi Wang, Jingechen Liu and Zhiliang Ying

2019 ICSA Applied Statistics Symposium, Raleigh, NC, June 9-12
Sandra Flores

Skew Testlet IRT Model under a Bayesian Approach
generalized linear model analysis of a single poset outcome. builds upon the author's previous published work which includes response model is implemented using existing software. This work parameters. Through a recoding scheme, the new multiple poset item component, it is also possible to glean information from the nominal nomination of weak ordering. Besides information from the ordinal com-
decomposes into separate ordinal and nominal models using the no-
be as high as p=60. In the talk I shall show that the poset modeling, the number of poset responses from the same individual could are collapsed into one category. The talk focuses on recent work and analyzed as ordinal or categorical data. For example, B and C because of the lack of analytic tools, poset is often first preprocessed cause of the lack of analytic tools, poset is often first preprocessed and analyzed as ordinal or categorical data. For example, B and C are collapsed into one category. The talk focuses on recent work on the theory and application of a latent variable model for multiple poset item responses. In some situations such as in educational testing, the number of poset responses from the same individual could be as high as p=60. In the talk I shall show that the poset model decomposes into separate ordinal and nominal models using the notion of weak ordering. Besides information from the ordinal component, it is also possible to glean information from the nominal component, which can lead to improved estimates of personal parameters. Through a recoding scheme, the new multiple poset item response model is implemented using existing software. This work builds upon the author’s previous published work which includes generalized linear model analysis of a single poset outcome.

A new latent variable model for the analysis of multiple partially ordered responses
Edward Ip
Wake Forest University
pr@wakehealth.edu

Partially ordered set (poset) responses form a class of data that is “in between” ordinal and categorical responses. One example of poset arises from response categories from a survey question about smoking: (A) never smoked, (B) former smoker, (C) current nondaily smoker, and (D) current daily smoker. The categories cannot be linearly rank-ordered, but a partial order does exist: A > B, C > D, in which categories B and C are not directly comparable. Poset response is actually quite common in the real world. However, because of the lack of analytic tools, poset is often first preprocessed and analyzed as ordinal or categorical data. For example, B and C are collapsed into one category. The talk focuses on recent work on the theory and application of a latent variable model for multiple poset item responses. In some situations such as in educational testing, the number of poset responses from the same individual could be as high as p=60. In the talk I shall show that the poset model decomposes into separate ordinal and nominal models using the notion of weak ordering. Besides information from the ordinal component, it is also possible to glean information from the nominal component, which can lead to improved estimates of personal parameters. Through a recoding scheme, the new multiple poset item response model is implemented using existing software. This work builds upon the author’s previous published work which includes generalized linear model analysis of a single poset outcome.

Skew Testlet IRT Model under a Bayesian Approach
Sandra Flores1, Caio Azevedo2, Jorge L Bazán3 and Dipak Dey4
1University of Sao Paulo
2University of Campinas
3University of Sao Paulo
4University of Connecticut
jlbazan@icmc.usp.br

In this work we introduce a new class of item response models that considered, simultaneously both testlet structures and asymmetric shape of Item Characteristic Curves (ICCs). Specifically, we have developed a skew testlet item response model, combining the random effects structure of testlet models and the logistic positive exponent item characteristic curve. Bayesian inference is developed through MCMC algorithms, including parameter estimation, model comparison and modelfit assessment. To detect the asymmetric behavior of the ICCs we proposed two procedures: one based on the marginal item information function and another through a mixture structure for the related prior distribution. The simulation studies indicated that the parameters are well recovered and that both testlet structure and asymmetric behavior of the ICCs are properly detected. Finally, a real data set, with testlet structures and empirical evidence of asymmetric behavior of the ICCs, were analyzed where it is shown that the proposed model captures the testlet effect and the asymmetric behavior of the ICC.

Session 100: How Data Science Drives Success in the Enterprises

Journey to an AI future
Bo Zhang
IBM
bozhang@us.ibm.com

When it comes to data science and AI, approaches of the past will not unlock the business value which companies need to sustain competitive advantage today, or in the future. This is why IBM introduced the AI Ladder, all the steps companies need to take to get to an AI future. In this talk, we’ll talk about the approaches and tools needed to accelerate this journey to AI, stretching across data & information architecture (IA) to analytics to machine learning to AI.

Deploy and monitor machine learning models in production
Ke Wei Wang, Ke Zhu, Matthew Neal and Emma Dickson
IBM
liwei.Wang@ibm.com

Nowadays almost all industries employ machine learning models to help on all kinds of tasks, e.g. recommendations, classifications, predictions, and etc. Data scientists may start with toy datasets and try different machine learning models. Then they will face the requirement on training a production-ready model on full data, deploying the model and maintaining the model’s performance on the go. IBM Watson Machine Learning provides the pipeline of services for both statisticians and non-statisticians from building a trustworthy machine learning model to deploying the trained model, and to continuous monitoring on the model’s health in a simplified way.

A progressive journey of applying data science in client advocacy
Dan Yang
IBM
YANGDA@US.IBM.COM

IBM Technical Support handles millions of cases per year and is committed to helping clients solve technical issues and creating optimal client experience. An always-on prediction system was developed to automatically forecast the daily client experience outcomes at the case level prior to case resolution. IBM Support teams use it to prioritize their workload and improve client experience of high risk cases. In this session, you’ll hear more about the progressive journey of developing the prediction system by combining classical statistical methods, machine learning algorithms, and IBM’s Watson Studio platform via iterations.

build an open source data lake for data scientists
Ke Zhu, Matthew Neal, Emma Dickson, Liwei Wang and Justin Eyster
IBM
kzhu@us.ibm.com

Data lake becomes a popular system and practice for data science projects within enterprise especially data driven companies.
talk will cover a number of usages of data lake and why our data scientists like it. this data lake design evolves over time from a combination of single vendor cloud services to cloud native, open source based technology by following a few design constraints without compromising performance nor scalability. It also covers lesson learned about using data lake in collaboration between data engineers and data scientists, including separate data store and computing, change management for catalog and schema and data access controls.

Session 101: New development in statistical methodologies for analyzing big neuroimaging data

Nonparametric Matrix Response Regression with Application to Brain Imaging Data Analysis
Wei Hu\(^1\), Dehan Kong\(^2\) and Weining Shen\(^1\)
\(^1\)University of California, Irvine
\(^2\)University of Toronto

With the rapid growth of neuroimaging technologies, a great effort has been dedicated recently to investigate the dynamic changes in brain activity. Examples include time course calcium imaging and dynamic brain functional connectivity. In this paper, we propose a novel nonparametric matrix response regression model to characterize the association between 2D image outcomes and predictors such as time and patient information. Our estimation procedure can be formulated as a nuclear norm regularization problem, which can capture the underlying low-rank structures of the dynamic 2D images. We develop an efficient algorithm to solve the optimization problem and introduce a Bayesian information criterion for our model to select the tuning parameters. Asymptotic theories including the risk bound and rank consistency are derived. We finally evaluate the empirical performance of our method using numerical simulations and real data applications from a calcium imaging study and an electroencephalography study.

Functional Regression Analysis of Distributional Data using Quantile Functions
Hojin Yang\(^1\), Veerabhadran Baladandayuthapani\(^2\) and Jeffrey Morris\(^1\)
\(^1\)University of Texas MD Anderson Cancer Center
\(^2\)University of Michigan

The aims of this paper are to look at the subject-specific distribution from observing the large number of repeated measurements for each subject and to determine how a set of covariates effects various aspects of the underlying subject-specific distribution, including the mean, median, variance, skewness, heavy-tailedness, and various upper and lower quantiles. To address these, we develop a quantile functional regression modeling framework that models the distribution of a set of common repeated observations from a subject through the quantile function. To account for smoothness in the quantile functions, we introduce novel basis functions adapting to the features of a given data set. Then, we build a Bayesian framework that uses nonlinear shrinkage of basis coefficients to regularize the functional regression coefficients and allows fully Bayesian inferences after fitting a Markov chain Monte Carlo. We demonstrate the benefit of the basis space modeling through simulation studies, and illustrate the method using a biomedical imaging data set in which we relate the distribution of pixel intensities from a tumor image to various demographic, clinical, and genetic characteristics.

Zero-Inflated Regime-Switching Stochastic Differential Equation Models for Highly Unbalanced Multiva
Zhaohua Lu\(^1\), Sy-Min Chun\(^2\), Nilam Ram\(^2\) and Pamela M. Cole\(^2\)
\(^1\)St. Jude Children’s Research Hospital
\(^2\)Pennsylvania State University

In the study of human dynamics, the behavior under study is often operationalized by tallying the frequencies and intensities of a collection of lower-order processes. However, because of idiosyncratic differences in how negative affect is expressed, some of the lower-order processes may be characterized by sparse occurrences in some individuals. We propose adding a regime (unobserved state) of “nonoccurrence” to a bivariate Ornstein-Uhlenbeck (OU) model to account for the high instances of non-occurrence in some individuals while simultaneously allowing for multivariate dynamic representation of the processes of interest under non-zero responses. The transition between the occurrence (i.e., active) and non-occurrence (i.e., inactive) regimes is represented using a latent Markovian transition model with dependencies on latent variables and person-specific covariates to account for inter-individual heterogeneity of the processes. Bayesian estimation and inference are based on Markov chain Monte Carlo algorithms. We demonstrate the utility of the proposed zero-inflated regime-switching OU model to a study of young children’s self-regulation at 36 and 48 months.

Dimension deduction method for group analysis of functional neuroimaging data
MiHye Ahn
University of Nevada-Reno

Recently, much attention has been received on the analysis of functional imaging data to delineate the intrinsic functional connectivity pattern among different brain regions within each subject. However, only few approaches for integrating functional connectivity pattern from multiple subjects have been proposed. The goal of this study is to develop a reduced-rank model framework for analyzing the whole-brain voxel-wise functional images across multiple subjects in the frequency domain. Considering the neighboring voxels with different weights, the frequency and spatial factors can be extracted. Imposing sparsity on the frequency factors enables us to identify the dominant frequencies. In addition, the spatial maps can be used for detecting group difference, when the comparison between different groups is of specific interest. Simulation study shows that the proposed method achieves less spatial variability and better estimates of frequency and spatial factors, compared to some existing methods. Finally, we apply the proposed method to ADNI data.

Session 102: Contributed Session 3

Sparse SIR: Optimal Rates and Adaptive Estimation
Kai Tan\(^1\), Lei Shi\(^2\) and Zhou Yu\(^3\)
\(^1\)University of Kentucky
\(^2\)Fudan University
\(^3\)East China Normal University

Sliced inverse regression (SIR) is an innovative and effective method for sufficient dimension reduction and data visualization. Recently, an impressive range of penalized SIR methods has been proposed to estimate the central subspace in a sparse fashion. Nonetheless, few of them considered the sparse sufficient dimension reduction from a decision-theoretic point of view. To address
this issue, we in this paper establish the minimax rates of convergence for estimating the sparse SIR directions under various commonly used loss functions in the literature of sufficient dimension reduction. We also discover the possible trade-off between statistical guarantee and computational performance for sparse SIR. We finally propose an adaptive estimation scheme for sparse SIR which is computationally tractable and rate optimal. Numerical studies are carried out to confirm the theoretical properties of our proposed methods.

Robust Moderately Clipped LASSO for Simultaneous Outlier Detection and Variable Selection

Yang Peng and Xiaoli Gao
University of North Carolina at Greensboro
yangp2019@gmail.com

Outlier detection has become an important and challenging issue in High-dimensional data analysis due to the coexistence of data contamination and high-dimensionality. Most existing widely used penalized least squares methods are sensitive to outliers due to the l2 loss. In this paper, we propose a Robust Moderately Clipped LASSO (RMCL) estimator, that performs simultaneous outlier detection, variable selection and robust estimation. The RMCL estimator can be efficiently solved using the coordinate descent algorithm in a convex-concave procedure. Our numerical studies demonstrate that the RMCL estimator possesses superiority in both variable selection and outlier detection and thus can be advantageous in difficult prediction problems with data contamination.

Online Updating of Information Based Model Selection Criterion in the Big Data Setting

Yishu Xue and Guanyu Hu
University of Connecticut
yishu.xue@uconn.edu

The generalized information criterion (GIC) is an important tool for model selection in statistical inference. In the big data setting, traditional GIC can not be calculated when the data size exceeds the computer memory. We propose an online updating approach to calculate the GIC and perform model selection for huge datasets. Specifically, we define the online updating versions of GICs for streaming data for the normal linear regression and generalized linear models. Under reasonable regularity conditions, we show that the information criterion selection procedures are asymptotically valid. The performance of the proposed criteria is assessed using extensive simulation study. The usage of our proposed model selection procedure is further illustrated with the analysis of two large datasets, the cover type data and the earthquake data. For both datasets, the online updating procedure selected the same or similar model as the entire data based model selection procedure.

Integrative Factorization of Bidimensionally Linked Matrices

Jun Young Park and Eric Lock
University of Minnesota
park1131@umn.edu

Advances in molecular "omics" technologies have motivated new methodology for the integration of multiple sources of high-content biomedical data. However, most statistical methods for integrating multiple data matrices only consider data shared vertically (one cohort on multiple platforms) or horizontally (different cohorts on a single platform). This is limiting for data that take the form of bidimensionally linked matrices (e.g., multiple cohorts measured on multiple platforms), which are increasingly common in large-scale biomedical studies. We propose BIDIFAC (Bidimensional Integrative Factorization) for integrative dimension reduction and signal approximation of bidimensionally linked data matrices. Our method factorizes the data into (i) globally shared, (ii) row-shared, (iii) column-shared, and (iv) single-matrix structural components, facilitating the investigation of shared and unique patterns of variability. For estimation we use a penalized objective function that extends the nuclear norm penalization for a single matrix. As an alternative to the complicated rank selection problem, we use results from random matrix theory to choose tuning parameters. We apply our method to integrate two genomics platforms (mRNA and miRNA expression) across two sample cohorts (tumor samples and normal tissue samples) using the breast cancer data from TCGA.

Session 103: Modern Statistical Methods for Complex Data

CHAMP: Post-Processing Partitions for Modularity Optimization

Peter Mucha
University of North Carolina at Chapel Hill
mucha@unc.edu

We present the Convex Hull of Admissible Modularity Partitions (CHAMP) algorithm, pruning and prioritizing different network community structures identified across multiple runs of possibly different computational heuristics. Given a set of partitions, CHAMP identifies the domain of modularity optimization for each partition – i.e., the parameter-space domain where it has the largest modularity relative to the input set – discarding partitions with empty domains to obtain the subset of partitions that are “admissible” candidate community structures that remain potentially optimal over indicated parameter domains. Importantly, CHAMP can be used for multi-dimensional parameter spaces, such as those for multilayer networks where one includes a resolution parameter and interlayer coupling. Using the results from CHAMP, a user can more appropriately select robust community structures by observing the sizes of domains of optimization and the pairwise comparisons between partitions in the admissible subset. We demonstrate the utility of CHAMP with several example networks, obtaining subsets of admissible partitions that are 20-to-1785 times smaller than the sets of unique partitions that were input into CHAMP.

Estimating Plant Growth Curves and Derivatives by Modeling Crowdsourced Imaged-based Data

Haozhe Zhang1, Dan Nettleton1, Stefan Hey1, Talukder Jubery1, Cheng-Ting Yeh1, Zihao Zheng1 and Pat Schnable3

1Iowa State University
dnett@iastate.edu

Recent advances in field-based plant phenotyping have increased interest in statistical approaches for analysis of longitudinal phenotypic data derived from sequential images. In a maize growth study, plants of various genotypes were imaged daily during the growing season by hundreds of cameras. Amazon Mechanical Turk (MTurk) workers were hired to manually mark plant bodies on these images, from which plant heights were obtained. An important scientific problem is to estimate the effect of genotype and its interaction with environment on plant growth while adjusting for measurement errors from crowdsourced image analysis. We model plant height measurements as discrete observations of latent smooth growth curves contaminated with MTurk worker random effects and worker-specific measurement errors. We allow the mean function of the growth curve and its first derivative to depend on replicates and environmental conditions, and model the phenotypic variation...
between genotypes and genotype-by-environment interactions by functional random effects. We estimate the covariance functions of the functional random effects by a fast penalized tensor product spline approach, and then perform robust functional principal component analysis (rFPCA) using the best linear unbiased predictor of the principal component scores. As byproducts, the proposed model leads to a new method for assessing the quality of MTurk worker data and a novel index for measuring the sensitivity to drought for various genotypes. The properties and advantages of the proposed approach are demonstrated by simulation studies.

Simultaneous Confidence Corridors for Mean Functions in Functional Data Analysis of Imaging Data

*Yueying Wang*, Guannan Wang, Li Wang and R. Todd Ogden

1 Iowa State University
2 College of William & Mary
3 Columbia University

yueyingw@iastate.edu

Motivated by recent work involving the analysis of biomedical imaging data, we present a novel procedure for constructing simultaneous confidence corridors for the mean of imaging data. We propose to use flexible bivariate splines over triangulations to handle irregular domain of the images that is common in brain imaging studies and in other biomedical imaging applications. The proposed spline estimators of the mean functions are shown to be consistent and asymptotically normal under some regularity conditions. We also provide a computationally efficient estimator of the covariance function and derive its uniform consistency. The procedure is also extended to the two-sample case in which we focus on comparing the mean functions from two populations of imaging data. Through Monte Carlo simulation studies we examine the finite-sample performance of the proposed method. Finally, the proposed method is applied to analyze brain Positron Emission Tomography (PET) data in two different studies. One dataset used in preparation of this article was obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database.

An adaptive estimation in dimension reduction

Qin Wang

University of Alabama
qwang57@ua.edu

Sufficient dimension reduction is a useful tool to extract the regression information through a small set of linear combinations of the original predictors. Motivated by a recent study where the central subspace was explored through conditional quantiles, we propose to use the regression expectiles to facilitate the adaptive estimation. An efficient algorithm will be discussed, and numerical examples will be presented.
Index of Authors
<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang, HJ</td>
<td>52</td>
<td>108</td>
</tr>
<tr>
<td>Wang, J</td>
<td>55</td>
<td>120</td>
</tr>
<tr>
<td>Wang, W</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>Wang, L</td>
<td>56</td>
<td>57</td>
</tr>
<tr>
<td>Wu, M</td>
<td>59</td>
<td>132</td>
</tr>
<tr>
<td>Wu, C</td>
<td>59</td>
<td>135</td>
</tr>
<tr>
<td>Wong, KY</td>
<td>56</td>
<td>57</td>
</tr>
<tr>
<td>Wu, J</td>
<td>59</td>
<td>132</td>
</tr>
<tr>
<td>White, R</td>
<td>56</td>
<td>57</td>
</tr>
<tr>
<td>Wei, Y</td>
<td>59</td>
<td>132</td>
</tr>
<tr>
<td>Wang, J</td>
<td>55</td>
<td>119</td>
</tr>
<tr>
<td>Yin, X</td>
<td>58</td>
<td>130</td>
</tr>
<tr>
<td>Ying, Z</td>
<td>60</td>
<td>136</td>
</tr>
<tr>
<td>Yu, B</td>
<td>49</td>
<td>98</td>
</tr>
<tr>
<td>Yu, G</td>
<td>56</td>
<td>124</td>
</tr>
<tr>
<td>Yu, L</td>
<td>42</td>
<td>58</td>
</tr>
<tr>
<td>Yu, M</td>
<td>51</td>
<td>105</td>
</tr>
<tr>
<td>Yu, S</td>
<td>45</td>
<td>80</td>
</tr>
<tr>
<td>Yu, T</td>
<td>45</td>
<td>82</td>
</tr>
<tr>
<td>Yu, W</td>
<td>42</td>
<td>70</td>
</tr>
<tr>
<td>Yuan, C</td>
<td>59</td>
<td>133</td>
</tr>
<tr>
<td>Yuan, M</td>
<td>37</td>
<td>87</td>
</tr>
<tr>
<td>Yuan, Q</td>
<td>58</td>
<td>130</td>
</tr>
<tr>
<td>Yue, L</td>
<td>40</td>
<td>49</td>
</tr>
<tr>
<td>Yin, J</td>
<td>55</td>
<td>119</td>
</tr>
<tr>
<td>Zhang, X</td>
<td>50</td>
<td>88</td>
</tr>
<tr>
<td>Zhao, H</td>
<td>48</td>
<td>91</td>
</tr>
<tr>
<td>Zhao, N</td>
<td>57</td>
<td>113</td>
</tr>
<tr>
<td>Zhao, Q</td>
<td>54</td>
<td>113</td>
</tr>
<tr>
<td>Zhao, S</td>
<td>59</td>
<td>133</td>
</tr>
<tr>
<td>Zhao, Y</td>
<td>51</td>
<td>55</td>
</tr>
<tr>
<td>Wang, X</td>
<td>42</td>
<td>97</td>
</tr>
<tr>
<td>Wang, Y</td>
<td>43</td>
<td>97</td>
</tr>
<tr>
<td>Wang, T</td>
<td>43</td>
<td>97</td>
</tr>
<tr>
<td>Wang, Z</td>
<td>43</td>
<td>97</td>
</tr>
<tr>
<td>Warren, J</td>
<td>50</td>
<td>101</td>
</tr>
<tr>
<td>Wei, G</td>
<td>41</td>
<td>67</td>
</tr>
<tr>
<td>Wei, J</td>
<td>59</td>
<td>135</td>
</tr>
<tr>
<td>Wei, P</td>
<td>55</td>
<td>117</td>
</tr>
<tr>
<td>Wei, Y</td>
<td>54</td>
<td>116</td>
</tr>
<tr>
<td>Weiss, H</td>
<td>59</td>
<td>135</td>
</tr>
<tr>
<td>Weng, H</td>
<td>49</td>
<td>96</td>
</tr>
<tr>
<td>Weng, J</td>
<td>58</td>
<td>130</td>
</tr>
<tr>
<td>West, BT</td>
<td>40</td>
<td>65</td>
</tr>
<tr>
<td>White, R</td>
<td>44</td>
<td>78</td>
</tr>
<tr>
<td>Wong, KY</td>
<td>45</td>
<td>73</td>
</tr>
<tr>
<td>Wong, RKW</td>
<td>57</td>
<td>127</td>
</tr>
<tr>
<td>Wu, C</td>
<td>59</td>
<td>133</td>
</tr>
<tr>
<td>Wu, CO</td>
<td>60</td>
<td>135</td>
</tr>
<tr>
<td>Wu, H</td>
<td>52</td>
<td>108</td>
</tr>
<tr>
<td>Wu, L</td>
<td>59</td>
<td>135</td>
</tr>
<tr>
<td>Wu, M</td>
<td>53</td>
<td>111</td>
</tr>
<tr>
<td>Wu, P</td>
<td>43</td>
<td>72</td>
</tr>
<tr>
<td>Wu, Q</td>
<td>44</td>
<td>76</td>
</tr>
<tr>
<td>Wu, TT</td>
<td>41</td>
<td>67</td>
</tr>
<tr>
<td>Wu, W</td>
<td>40</td>
<td>63</td>
</tr>
<tr>
<td>Wu, Y</td>
<td>44</td>
<td>50</td>
</tr>
<tr>
<td>Wu, Z</td>
<td>58</td>
<td>129</td>
</tr>
<tr>
<td>Xu, D</td>
<td>51</td>
<td>59</td>
</tr>
<tr>
<td>Xiao, F</td>
<td>42</td>
<td>70</td>
</tr>
<tr>
<td>Xiao, L</td>
<td>41</td>
<td>68</td>
</tr>
<tr>
<td>Xie, H</td>
<td>59</td>
<td>133</td>
</tr>
<tr>
<td>Xie, S</td>
<td>48</td>
<td>54</td>
</tr>
<tr>
<td>Xie, Y</td>
<td>56</td>
<td>123</td>
</tr>
<tr>
<td>Xiong, F</td>
<td>54</td>
<td>113</td>
</tr>
<tr>
<td>Xiong, X</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>Xu, G</td>
<td>41</td>
<td>58</td>
</tr>
<tr>
<td>Xu, H</td>
<td>57</td>
<td>126</td>
</tr>
<tr>
<td>Xu, J</td>
<td>49</td>
<td>97</td>
</tr>
<tr>
<td>Xu, L</td>
<td>53</td>
<td>113</td>
</tr>
<tr>
<td>Xu, Y</td>
<td>44</td>
<td>47</td>
</tr>
<tr>
<td>Xue, F</td>
<td>60</td>
<td>136</td>
</tr>
<tr>
<td>Xue, L</td>
<td>46</td>
<td>83</td>
</tr>
<tr>
<td>Xue, X</td>
<td>50</td>
<td>101</td>
</tr>
<tr>
<td>Xue, Y</td>
<td>61</td>
<td>139</td>
</tr>
<tr>
<td>Yan, B</td>
<td>52</td>
<td>106</td>
</tr>
<tr>
<td>Yan, D</td>
<td>59</td>
<td>135</td>
</tr>
<tr>
<td>Yan, X</td>
<td>50</td>
<td>63</td>
</tr>
<tr>
<td>Yang, D</td>
<td>60</td>
<td>137</td>
</tr>
<tr>
<td>Yang, F</td>
<td>52</td>
<td>52</td>
</tr>
<tr>
<td>Yang, H</td>
<td>60</td>
<td>138</td>
</tr>
<tr>
<td>Yang, J</td>
<td>41</td>
<td>51</td>
</tr>
<tr>
<td>Yang, L</td>
<td>45</td>
<td>80</td>
</tr>
<tr>
<td>Yang, S</td>
<td>42</td>
<td>43</td>
</tr>
<tr>
<td>Yang, T</td>
<td>60</td>
<td>137</td>
</tr>
<tr>
<td>Yang, Z</td>
<td>44</td>
<td>77</td>
</tr>
<tr>
<td>Yao, X</td>
<td>47</td>
<td>88</td>
</tr>
<tr>
<td>Ye, J</td>
<td>53</td>
<td>109</td>
</tr>
<tr>
<td>Ye, Y</td>
<td>53</td>
<td>113</td>
</tr>
<tr>
<td>Yeh, C</td>
<td>41</td>
<td>139</td>
</tr>
<tr>
<td>Yi, G</td>
<td>43</td>
<td>81</td>
</tr>
<tr>
<td>Zou, F</td>
<td>48</td>
<td>91</td>
</tr>
<tr>
<td>Zhuo, B</td>
<td>41</td>
<td>65</td>
</tr>
<tr>
<td>Zhang, W</td>
<td>43</td>
<td>74</td>
</tr>
<tr>
<td>Zhang, X</td>
<td>43</td>
<td>74</td>
</tr>
<tr>
<td>Zhang, Y</td>
<td>55</td>
<td>120</td>
</tr>
<tr>
<td>Zeleniuch-Jacquotte, A</td>
<td>48</td>
<td>94</td>
</tr>
<tr>
<td>Zeng, D</td>
<td>32</td>
<td>43</td>
</tr>
<tr>
<td>Zeng, F</td>
<td>54</td>
<td>69</td>
</tr>
<tr>
<td>Zhan, Q</td>
<td>48</td>
<td>91</td>
</tr>
<tr>
<td>Zhan, X</td>
<td>52</td>
<td>52</td>
</tr>
<tr>
<td>Zhang, B</td>
<td>40</td>
<td>49</td>
</tr>
<tr>
<td>Zhang, C</td>
<td>55</td>
<td>120</td>
</tr>
<tr>
<td>Zhang, D</td>
<td>55</td>
<td>119</td>
</tr>
<tr>
<td>Zhang, H</td>
<td>58</td>
<td>51</td>
</tr>
<tr>
<td>Zhang, J</td>
<td>22</td>
<td>51</td>
</tr>
<tr>
<td>Zhang, K</td>
<td>46</td>
<td>51</td>
</tr>
<tr>
<td>Zhang, L</td>
<td>49</td>
<td>55</td>
</tr>
<tr>
<td>Zhang, S</td>
<td>49</td>
<td>50</td>
</tr>
<tr>
<td>Zhang, W</td>
<td>43</td>
<td>44</td>
</tr>
<tr>
<td>Zhong, X</td>
<td>54</td>
<td>57</td>
</tr>
<tr>
<td>Zhou, Y</td>
<td>48</td>
<td>77</td>
</tr>
<tr>
<td>Zhou, Z</td>
<td>55</td>
<td>120</td>
</tr>
<tr>
<td>Zhou, S</td>
<td>59</td>
<td>135</td>
</tr>
<tr>
<td>Zhou, W</td>
<td>44</td>
<td>51</td>
</tr>
<tr>
<td>Zhou, Y</td>
<td>42</td>
<td>65</td>
</tr>
<tr>
<td>Zhou, B</td>
<td>41</td>
<td>65</td>
</tr>
<tr>
<td>Zhou, H</td>
<td>47</td>
<td>49</td>
</tr>
<tr>
<td>Zhu, J</td>
<td>53</td>
<td>111</td>
</tr>
<tr>
<td>Zhu, K</td>
<td>60</td>
<td>137</td>
</tr>
<tr>
<td>Zhu, L</td>
<td>56</td>
<td>124</td>
</tr>
<tr>
<td>Zhu, R</td>
<td>56</td>
<td>127</td>
</tr>
<tr>
<td>Zhou, W</td>
<td>44</td>
<td>51</td>
</tr>
<tr>
<td>Zhu, Y</td>
<td>42</td>
<td>45</td>
</tr>
<tr>
<td>Zhuo, B</td>
<td>41</td>
<td>65</td>
</tr>
<tr>
<td>Zhu, H</td>
<td>53</td>
<td>87</td>
</tr>
<tr>
<td>Zhu, N</td>
<td>112</td>
<td></td>
</tr>
<tr>
<td>Zou, J</td>
<td>49</td>
<td>96</td>
</tr>
</tbody>
</table>
See you in Houston, Texas in 2020

www.icsa.org